PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : H04N 7/30

A1

(11) International Publication Number: WO 98/38800

(43) International Publication Date: 3 September 1998 (03.09.98)

(21) International Application Number: PCT/GB98/00582

(22) International Filing Date: 25 February 1998 (25.02.98)

(30) Priority Data:

9703834.3	25 February 1997 (25.02.97)	GB
9703831.9	25 February 1997 (25.02.97)	GB

(71) Applicants *(for all designated States except US)*: BRITISH BROADCASTING CORPORATION [GB/GB]; Broadcasting House, London W1A 1AA (GB). SNELL & WILCOX LIMITED [GB/GB]; 6 Old Lodge Place, St. Margaret's, Twickenham, Middlesex TW1 1RQ (GB).

(72) Inventors; and

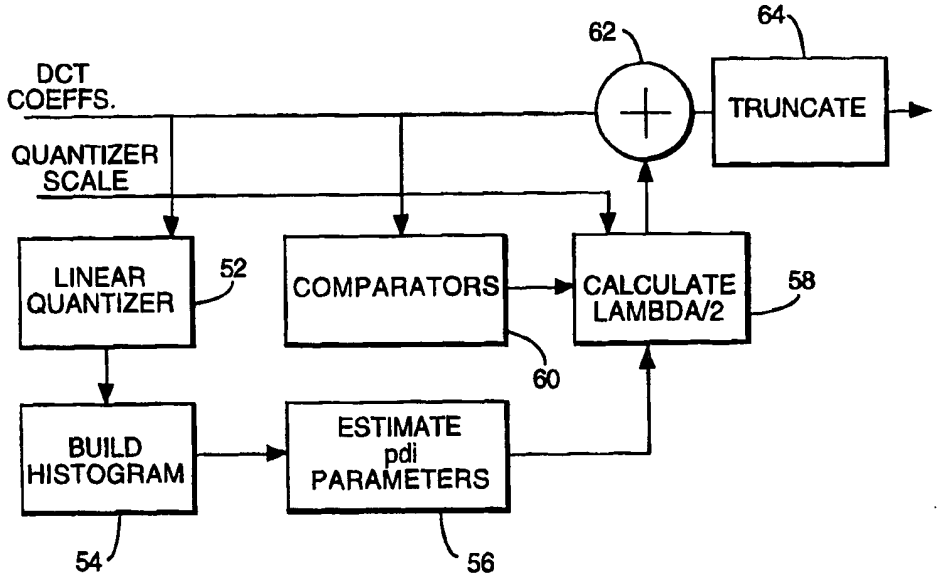
(75) Inventors/Applicants *(for US only)*: WERNER, Oliver, Hartwig [DE/GB]; Flat 1, Birdhurst Road 14 C, South Croydon, Surrey CR2 7EA (GB). WELLS, Nicholas, Dominic [GB/GB]; 17 Wellington Road, Brighton, East Sussex BN2 2AB (GB). KNEE, Michael, James [GB/GB]; 6 Woodbury Avenue, Petersfield, Hampshire GU32 2EE (GB).

(74) Agent: GARRATT, Peter, Douglas; Mathys & Squire, 100 Gray's Inn Road, London WC1X 8AL (GB).

(81) Designated States: AU, CA, JP, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: DIGITAL SIGNAL COMPRESSION ENCODING WITH IMPROVED QUANTISATION



(57) Abstract

In compression encoding of a digital signal, such as MPEG2, transform coefficients are quantised with the lower bound of each interval being controlled by a parameter λ. In the MPEG2 reference coder, for example, λ=0.75. Because the quantised coefficients are variable length coded, improved quality or reduced bit rates can be achieved by controlling λ so as to vary dynamically the bound of each interval with respect to the associated representation level. The parameter λ can vary with coefficient amplitude, with frequency, with quantisation step size. In a transcoding operation, λ can also vary with parameters in the initial coding operation.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

DIGITAL SIGNAL COMPRESSION ENCODING WITH IMPROVED QUANTISATION

This invention relates to the compression of digital video, audio or other signals.

Compression encoding generally involves a number of separate techniques. These will usually include a transformation, such as the block-based discrete cosine transform (DCT) of MPEG-2; an optional prediction step; a quantisation step and variable length coding. This invention is particularly concerned in this context with quantisation.

The quantisation step maps a range of original amplitudes onto the same representation level. The quantisation process is therefore irreversible. MPEG-2, (in common with other compression standards such as MPEG-1, JPEG, CCITT/ITU-T Rec.H.261 and ITU-T Rec.H.263) defines representation levels and leaves undefined the manner in which the original amplitudes are mapped onto a given set of representation levels.

In general terms, a quantizer assigns to an input value, which may be continuous or may previously have been subjected to a quantisation process, a code usually selected from quantization levels immediately above and immediately below the input value. The error in such a quantization will generally be minimised if the quantization level closest to the input value is selected. In a compression system, it is further necessary to consider the efficiency with which respective quantization levels may be coded. In variable length coding, the quantization levels which are employed most frequently are assigned the shortest codes.

Typically, the zero level has the shortest code. A decision to assign a higher quantization level, on the basis that it is the closest, rather than a lower level (and especially the zero level) will therefore decrease coding efficiency. In MPEG2, the overall bit rate of the compressed signal is maintained beneath a pre-determined limit by increasing the separation of quantization levels in response to a tendency toward higher bit rate. Repeated decisions to assign quantization levels on the basis of which is closest, may through

coding inefficiency thus lead to a coarser quantization process.

The behaviour of a quantizer in this respect may be characterised through a parameter λ which is arithmetically combined with the input value, with one value of λ (typically $\lambda = 1$) representing the selection of the closest quantization level or "rounding". A different value of λ (typically $\lambda = 0$) will in contrast represent the automatic choice of the lower of the two nearest quantization levels, or "truncating". In the MPEG2 reference coder, an attempt is made to compromise between the nominal reduction in error which is the attribute of rounding and the tendency toward bit rate efficiency which is associated with truncating, by setting a standard value for λ of $\lambda = 0.75$.

Whilst particular attention has here been paid to MPEG2 coding, similar considerations apply to other methods of compression encoding of a digital signal, which including the steps of conducting a transformation process to generate values and quantising the values through partitioning the amplitude range of a value into a set of adjacent intervals, whereby each interval is mapped onto a respective one of a set of representation levels which are to be variable length coded, such that a bound of each interval is controlled by a parameter λ. The transformation process may take a large variety of forms, including block-based transforms such as the DCT of MPEG2, and sub-band coding.

It is an object of one aspect of the present invention to provide an improvement in such a method which enables higher quality to be achieved at a given bitrate or a reduction in bitrate for a given level of quality.

Accordingly, the present invention is in one aspect characterised in that λ is controlled so as to vary dynamically the bound of each interval with respect to the associated representation level.

Suitably, wherein each value is arithmetically combined with λ.

Advantageously, λ is :

a function of the quantity represented by the value;

where the transformation is a DCT, a function of horizontal and vertical frequency;

a function of the quantisation step size; or

a function of the amplitude of the value.

In a particular form of the present invention, the digital signal to be encoded has been subjected to previous encoding and decoding processes and λ is controlled as a function of a parameter in said previous encoding and decoding processes.

In a further aspect, the present invention consists in a (q, λ) quantiser operating on a set of transform coefficients x_k representative of respective frequency indices f_k in which λ is dynamically controlled in dependence upon the values of x_k and f_k.

Advantageously, λ is dynamically controlled to minimise a cost function $D + \mu H$ where D is a measure of the distortion introduced by the quantisation in the uncompressed domain and H is a measure of compressed bit rate.

The invention will now be described by way of example with reference to the accompanying drawings, in which:-

Figure 1 is a diagram illustrating the relationships between representation levels, decision levels and the value of λ;

Figure 2 is a block diagram representation of the quantization process in the MPEG2 reference coder;

Figure 3 is a block diagram representation of a simplified and improved quantization process;

Figure 4 is a block diagram representation of the core elements of Figure 3;

Figure 5 is a block diagram representation of a quantization process according to one aspect of the present invention; and

Figure 6 is a block diagram representation of a quantization process according to a further aspect of the present invention.

In the specifically mentioned compression standards, the original amplitude x results from a discrete cosine transform (DCT) and is thus related to a horizontal frequency index f_{hor} and a vertical frequency index f_{ver}. Whilst this approach is taken as an example in what follows, the invention is not restricted in this regard.

In general, a quantiser describes a mapping from an original amplitude x of frequencies f_{hor} and f_{ver} onto an amplitude $y = Q(x)$. The mapping performed by the quantiser is fully determined by the set of representation levels $\{r_l\}$ and by the corresponding decision levels $\{d_l\}$ as illustrated in Figure 1. All original amplitudes in the range $d_l \le x < d_{(l+1)}$ are mapped onto the same representation level $y = Q(x) = r_l$. As can be seen from Figure 1, consecutive decision levels are related by the quantisation step size q: and for a given representation level r_l, the corresponding decision level is

$$d_{l-1} = d_l + q \qquad (1)$$

calculated as:

$$d_l = r_l - \frac{\lambda}{2} \cdot q \qquad (2)$$

The quantiser is fully specified by the quantisation step-size q and the parameter λ for a given set of representation levels $\{r_l\}$. Therefore, a quantiser that complies with equations (1) and (2) can be referred to as a (q,λ) quantiser.

Currently proposed quantisers, as described in the reference coders for the H.261, H.263, MPEG-1 and MPEG-2 standards, all apply a special type of (q, λ) quantiser in that a fixed value of λ is used: for example $\lambda = 0.75$ in the MPEG-2 reference coder or $\lambda = 1.0$ in the MPEG-1 reference coder for quantisation of intra-DCT-coefficients.

According to one aspect of this invention, λ is not constant but is a function that depends on the horizontal frequency index f_{hor}, the vertical

frequency index f_{ver}, the quantisation step-size q and the amplitude x:

$$\lambda = \lambda(f_{hor}, f_{ver}, q, x) \quad (3)$$

Examples of ways in which the function may usefully be derived to improve picture quality in video compression at a given bit-rate - or to reduce the required bit-rate at a given picture quality - will be set out below.

The invention extends also to the case of transcoding when a first generation amplitude $y_1 = Q_1(x)$ is mapped onto a second generation amplitude $y_2 = Q_2(y_1)$ to further reduce the bit-rate from the first to the second generation without having access to the original amplitude x. In this case the first generation quantiser Q_1 and the second generation quantiser Q_2 are described as a (q_1, λ_1)-type quantiser and a (q_2, λ_2)-type quantiser, respectively. The second generation λ_2 value is described as a function:

$$\lambda_2 = \lambda_2(f_{hor}, f_{ver}, q_1, \lambda_1, q_2, \lambda_{2,ref}, y_1) \quad (4)$$

The parameter $\lambda_{2,ref}$ that appears in Eqn. (4) is applied in a reference $(q_2, \lambda_{2,ref})$-type quantiser. This reference quantiser bypasses the first generation and directly maps an original amplitude x onto a second generation reference amplitude $y_{2,ref} = Q_{2,ref}(x)$.

The functional relationship of Eqn. (4) can be used to minimise the error $(y_2 - y_{2,ref})$ or the error $(y_2 - x)$. In the first case, the resulting second generation quantiser may be called a maximum *a-posteriori* (MAP) quantiser. In the second case, the resulting second generation quantiser may be called a mean squared error (MSE) quantiser. Examples of the second generation $(q_2, \lambda_{2,MAP})$-type and $(q_2, \lambda_{2,MSE})$-type quantisers are given below. For a more detailed explanation of the theoretical background, reference is directed to

the paper "Transcoding of MPEG-2 intra frames" - Oliver Werner - IEEE Transactions on Image Processing 1998, which will for ease of reference be referred to hereafter as "the Paper". A copy of the Paper is appended to British patent application No. 9703831 from which the present application claims priority.

The present invention refers specifically to quantization of 'intra' DCT coefficients in MPEG2 video coding but can be applied to non-intra coefficients, to other video compression schemes and to compression of signals other than video. In MPEG2, the prior art is provided by what is known as Test Model 5 (TM5). The quantization scheme of TM5 for positive intra coefficients is illustrated in Figure 2.

In order to simplify the description, the above diagram will be replaced by Figure 3, which illustrates essentially the same quantizer except for small values of **q**, where it corrects an anomaly as described in the Paper.

In this quantizer, the incoming coefficients are first divided by quantizer weighting matrix values, **W**, which depend on the coefficient frequency but which are fixed across the picture, and then by a quantizer scale value **q** which can vary from one macroblock to the next but which is the same for all coefficient frequencies. Prior to the adder, the equivalent inverse quantizer reconstruction levels are simply the integers 0, 1, 2 ... . A fixed number $\lambda/2$, is then added to the value and the result truncated. The significance of λ is that a value of 0 makes the quantizer (of the value input to the adder) a simple truncation, while a value of 1 makes it a rounding operation. In TM5, the value of λ is fixed at **0.75**.

Attention will hereafter be focused on the operation of the 'core' quantizer shown in Figure 4.

In a class of MPEG-2 compatible quantisers for,intra frame coding, non-negative original dct-coefficients ***x*** (or the same coefficents after division by weighting matrix values **W**) are mapped onto the representation levels as:

$$y = Q(x) = \left\lfloor \frac{x}{q} + \frac{\lambda}{2} \right\rfloor . q \qquad (5)$$

The floor function $\lfloor a \rfloor$ extracts the integer part of the given argument a.

Negative values are mirrored:

$$y = -Q(|x|) \qquad (6)$$

The amplitude range of the quantisation step-size q in eq. (1) is standardised; q has to be transmitted as side information in every MPEG-2 bit stream. This does not hold for the parameter λ in eq. (1). This parameter is not needed for reconstructing the dct-coefficients from the bit stream, and is therefore not transmitted. However, the λ-value controls the mapping of the original dct-coefficients x onto the given set of representation levels

$$r_l = l . q \qquad (7)$$

According to eq. (1), the (positive) x-axis is partitioned by the decision levels

$$d_l = \left(l - \frac{\lambda}{2} \right) . q \quad l = 1, 2, ... \qquad (8)$$

Each $x \in [\, d_l, d_{l+1})$ is mapped onto the representation level $y = r_l$. As a special case, the interval $[0, d_1)$ is mapped onto $y = 0$.

The parameter λ can be adjusted for each quantisation step-size q, resulting in a distortion rate optimised quantisation: the mean-squared-error

$$D = E[(x - y)^2] \qquad (9)$$

is minimised under a bit rate constraint imposed on the coefficients y. In order to simplify the analysis, the first order source entropy

$$H = \Sigma_l - P_l \cdot \log_2 P_l \qquad (10)$$

of the coefficients y instead of the MPEG-2 codeword table is taken to calculate the bit rate. It has been verified in the Paper that the entropy H can be used to derive a reliable estimate for the number of bits that result from the MPEG-2 codeword table. In Eqn. (10), P_l denotes the probability for the occurrence of the coefficient $y=r_l$.

The above constrained minimisation problem can be solved by applying the Lagrange multiplier method, introducing the Lagrange multiplier μ. One then gets the basic equation to calculate the quantisation parameter λ:

$$\frac{\partial D}{\partial \lambda} + \mu \cdot \frac{\partial H}{\partial \lambda} = 0 \qquad (11)$$

Note, that the solution for λ that one obtains from Eqn. (11) depends on the value of μ. The value of μ is determined by the bit rate constraint

$$H \leq H_0 \qquad (12)$$

where H_0 specifies the maximum allowed bit rate for encoding the coefficients y. In general, the amplitude range of the Lagrange multiplier is $0 < \mu < \infty$. In the special case of $H_0 \rightarrow \infty$, one obtains $\mu \rightarrow 0$. Conversely for $H_0 \rightarrow 0$, one obtains in general $\mu \rightarrow \infty$.

The Laplacian probability density function (pdf) is an appropriate model for describing the statistical distribution of the amplitudes of the original dct-coefficients. This model is now applied to evaluate analytically Eqn. (11). One then obtains a distortion-rate optimised quantiser characteristic by inserting the resulting value for λ in eq. (5).

Due to the symmetric quantiser characteristic for positive and negative

amplitudes in Eqns. (5) and (6), we introduce a pdf p for describing the distribution of the absolute original amplitudes $|x|$. The probability P_0 for the occurrence of the coefficient $y = 0$ can then be specified as

$$P_0 = \int_0^{\left(1-\frac{\lambda}{2}\right).q} p(x)dx \qquad (13)$$

Similarly, the probability P_l for the coefficient $|y|$ becomes

$$P_l = \int_{\left(l-\frac{\lambda}{2}\right).q}^{\left(l+1-\frac{\lambda}{2}\right).q} p(x)dx \quad l = 1,2, \ldots\ldots \qquad (14)$$

With Eqns. (13) and (14), the partial derivative of the entropy H of eq. (10) can be written after a straightforward calculation as

$$\frac{\partial H}{\partial \lambda} = \frac{q}{2} \cdot \sum_{l \geq 0} p\left(\left(l + 1 - \frac{\lambda}{2}\right).q\right) \cdot \log_2 \frac{P_l}{P_{l+1}} \qquad (15)$$

From eq. (9) one can first deduce

$$D = \int_0^{(1-\frac{\lambda}{2}).q} x^2 \cdot p(x)dx + \sum_{l \geq 1} \int_{(l-\frac{\lambda}{2}).q}^{(l+1-\frac{\lambda}{2}).q} (x - l.q)^2 \cdot p(x)dx \qquad (16)$$

and further from eq. (16)

$$\frac{\partial D}{\partial \lambda} = \frac{-q^3}{2} \cdot (1 - \lambda) \cdot \sum_{l \geq 0} p\left(\left(l + 1 - \frac{\lambda}{2}\right).q\right) \qquad (17)$$

It can be seen from eq. (17) that

$$\frac{\partial D}{\partial \lambda} \geq 0 \quad if\ 0 \leq \lambda \leq 1 \qquad (18)$$

Thus, when λ is increased from zero to one, the resulting distortion D is monotonically decreasing until the minimum value is reached for $\lambda = 1$. The latter is the solution to the unconstrained minimisation of the mean-squared-error, however, the resulting entropy H will in general not fulfil the bit rate constraint of eq. (12).

Under the assumption of $P_l \geq P_{l+1}$ in eq. (15), we see that $\partial H/\partial\lambda \geq 0$. Thus, there is a monotonic behaviour: when λ is increased from zero to one, the resulting distortion D monotonically decreases, at the same time the resulting entropy H montonically increases. Immediately, an iterative algorithm can be derived from this monotonic behaviour. The parameter λ is initially set to $\lambda = 1$, and the resulting entropy H is computed. If H is larger than the target bit rate H_0, the value of λ is decreased in further iteration steps until the bit rate constraint, eq. (12), is fulfilled. While this iterative procedure forms the basis of a simplified distortion-rate method proposed for transcoding of I-frames, we continue to derive an analytical solution for λ.

Eqns. (15) and (17) can be evaluated for the Laplacian model:

$$p(x) = \beta . \alpha . e^{-\alpha x} \quad if \quad x \geq d_l = \left(1 - \frac{\lambda}{2}\right).q \qquad (19)$$

After inserting the model pdf of Eqn. (19) in Eqns. (15) and (17), it can be shown that the basic equation (11) leads then to the analytical solution for λ,

$$\lambda = 1 - \frac{\mu}{q^2} . \left[h(z) + (1 - z) . \log_2\left(\frac{P_0}{1 - P_0}\right)\right] \qquad (20)$$

with $z = e^{-\alpha \cdot q}$ and the 'z'-entropy

$$h(z) = -z.\log_2 z - (1-z).\log_2(1-z) \qquad (21)$$

Eqn. (20) provides only an implicit solution for λ, as the probability P_0 on the right hand side depends on λ according to eq. (13). In general, the value of P_0 can be determined only for known λ by applying the quantiser characteristic of Eqns. (5) and (6) and counting the relative frequency of the event $y = 0$. However, eq. (20) is a fixed-point equation for λ which becomes more obvious if the right hand side is described by the function

$$g(\lambda) = 1 - \frac{\mu}{q^2}.\left[h(z) + (1-z).\log_2\left(\frac{P_0}{1-P_0}\right)\right] \qquad (22)$$

resulting in the classical fixed-point form λ = g(λ). Thus, it follows from the fixed point theorem of Stefan Banach that the solution for λcan be found by an iterative procedure with

$$\lambda_{j+1} = g(\lambda_j) \qquad (23)$$

in the (j + 1)-th iteration step. The iteration of (23) converges towards the solution for an arbitrary initial value λ_0 if the function *g* is 'self-contracting', i.e. Lipschitz-continuous with a Lipschitz-constant smaller than one. As an application of the mean theorem for the differential calculus, it is not difficult to prove that *g* is always 'self-contracting' if the absolute value of the partial derivative is less than one. This yields the convergence condition

$$1 > \left|\frac{\partial g}{\partial \lambda}\right| = \frac{1}{2.\ln(2)}.\frac{\mu}{q}.(1-z).\frac{\alpha}{P_0} \qquad (24)$$

A distortion-rate optimised quantisation method will now be derived based on the results obtained above. As an example, a technique is outlined for quantising the AC-coefficients of MPEG-2 intra frames. It is straightforward to modify this technique for quantising the dct-coefficients of MPEG-2 inter frames, i.e. P- and B-frames.

Firstly, one has to take into account that the 63 AC-coefficients of an 8x8 dct-block do not share the same distribution. Thus, an individual Laplacian model pdf according to eq. (19) with parameter α_i is assigned to each AC-frequency index *i*. This results in an individual quantiser characteristic according to Eqns. (5) and (6) with parameter λ_i. Furthermore, the quantisation step-size q_i depends on the visual weight w_i and a frequency-independent ***qscale*** parameter as

$$q_i = \frac{w_i \cdot qscale}{16} \qquad (25)$$

For a given step-size q_i, the quantisation results in a distortion $D_i(\lambda_i)$ and a bit rate $H_i(\lambda_i)$ for the AC-coefficients of the same frequency index *i*. As the dct is an orthogonal transform, and as the distortion is measured by the mean-squared-error, the resulting distortion *D* in the spatial (sample/pixel) domain can be written as

$$D = c \cdot \sum_i D_i(\lambda_i) \qquad (26)$$

with some positive normalising constant ***c***. Alternatively the distortion can measured in the weighted coefficient domain in order to compensate for the variation in the human visual response at different spatial frequencies.

Similarly, the total bit rate *H* becomes

$$H = \sum_i H_i\ (\lambda_i) \qquad (27)$$

For a distortion rate optimised quantisation, the 63 parameters λ_i have to be adjusted such that the cost function

$$D + \mu . H \qquad (28)$$

is minimised. The non-negative Lagrange multiplier **μ** is determined by the bit rate constraint

$$H \leq H_0 \qquad (29)$$

Alternatively, if the distortion is expressed in the logarithmic domain as:

$$D' = 20 \log_{10} D \ \text{dB} \qquad (28a)$$

The cost function to be minimised becomes:

$$B = D + \mu' H \qquad (28b)$$

Where μ' is now an *a priori* constant linking distortion to bit rate.

A theoretical argument based on coding white noise gives a law of **6 dB** per bit per coefficient. In practice, observation of actual coding results at different bit rates gives a law of **k dB** per bit, where **k** takes values from about 5 to about 8 depending on the overall bit rate. In practice, the intuitive '**6dB**' law corresponds well with observation.

Additionally, the *qscale* parameter can be changed to meet the bit rate constraint of Eqn. (25). In principle, the visual weights w_i offer another degree of freedom but for simplicity we assume a fixed weighting matrix as in the MPEG-2 reference decoder. This results in the following distortion rate optimised quantisation technique which can be stated in a 'C'-language-like form:

/* Begin of quantising the AC-coefficients in MPEG-2 intra frames*/

$D_{min} = \infty$;

for ($qscale = qmin$; $qscale \leq qmax$; $qscale = qscale + 2$)/* linear qscale table*/

{

$\mu = 0$;

do {

Step 1: determine $\lambda_1, \lambda_2, \ldots, \lambda_{63}$ by minimising $D + \mu \cdot H$;

Step 2: calculate $H = \Sigma H_i(\lambda_i)$;

$\mu = \mu + \delta$; /*δ to be selected appropriately*/

}while ($H > H_O$);

Step 3: calculate $D = c \cdot \Sigma D_i(\lambda_i)$;

if ($D < D_{min}$){

$qscale_{opt} = qscale$;

for ($i = 1$; $i \leq 63$; $i = i + 1$) $\lambda_{i,opt} = \lambda_i$;

$D_{min} = D$;}

}

for ($i = 1$; $i \leq 63$; $i = i + 1$)

$$q_{i,opt} = \frac{w_i \cdot qscale_{opt}}{16}$$

{

$$y = Q_i(x) = \left\lfloor \frac{|x|}{q_{i,opt}} + \frac{\lambda_{i,opt}}{2} \right\rfloor \cdot q_{i,opt} \cdot sgn(x)$$

quantise all AC-coefficients of frequency-index i by

}

/*End of quantising the AC-coefficients in MPEG-2 intra frames*/

There are several options for performing Step 1 - Step 3:

1. Options for performing Step 1

The parameters $\lambda_1, \lambda_2, ..., \lambda_{63}$ can be determined

a) analytically by applying Eqns. (20)-(23) of Section 3.

b) iteratively by dynamic programming of $D + \mu \cdot H$, where either of the options described in the next points can be used to calculate D and H.

2. Options for performing Step 2

$H = \Sigma H_i(\lambda_i)$ can be calculated

(a) by applying the Laplacian model pdf, resulting in

$$H = \sum_i h(P_{0,i}) + (1 - P_{0,i}) \cdot \frac{h(z_i)}{1 - z_i} \qquad (32)$$

where $h(P_{0,i})$ and $h(Z_i)$ are the entropies as defined in eq. (21) of P_{0i} (eq. (13)) and $Z_i = e^{-\alpha_i \cdot q_i}$, respectively. Note that $P_{0,i}$ in Eqn.(32) can be determined by counting for each dct-frequency index i the relative frequency of the zero-amplitude $y = Q_i(x) = 0$. Interestingly, eq. (32) shows that the impact of the quantisation parameters λ_i on the resulting bit rate H only consists in controlling the zero-amplitude probabilities $P_{0,i}$.

b) from a histogram of the original dct-coefficients, resulting with Eqns. (10), (13) and (14) in

$$H = -\sum_i \sum_l P_{l,i} \cdot \log_2 P_{l,i} \qquad (33)$$

c) by applying the MPEG-2 codeword table

3. Options for performing Step 3

$D = c \cdot \Sigma D_i(\lambda_i)$ can be calculated

a) by applying the Laplacian model pdf of Eqn. (19) and evaluating Eqn. (16).

b) by calculating $D = E[(x - y)^2]$ directly from a histogram of the original dct-coefficients x.

Depending on which options are chosen for Step 1 - Step 3, the proposed method results in a single pass encoding scheme if the Laplacian model pdf is chosen or in a multi pass scheme if the MPEG-2 codeword table is chosen. Furthermore, the method can be applied on a frame, macroblock or on a 8x8-block basis, and the options can be chosen appropriately. The latter is of particular interest for any rate control scheme that sets the target bit rate H_0 either locally on a macroblock basis or globally on a frame basis.

Furthermore, we note that the proposed method skips automatically high-frequency dct-coefficients if this is the best option in the rate-distortion sense. This is indicated if the final quantisation parameter $\lambda_{i,opt}$ has a value close to one for low-frequency indices i but a small value, e.g. zero, for high-frequency indices.

A distortion-rate optimised quantisation method for MPEG-2 compatible coding has been described, with several options for an implementation. The invention can immediately be applied to standalone (first generation) coding. In particular, the results help designing a sophisticated rate control scheme.

The quantiser characteristic of eqs. (5) and (6) can be generalised to

$$y = Q(x) = r(x) + \left\lfloor \frac{x - r(x)}{q(x)} + \frac{\lambda(x)}{2} \right\rfloor \cdot q(x) \qquad (34)$$

for non-negative amplitudes **x**. The floor-function $\lfloor a \rfloor$ in eq. (34) returns the integer part of the argument ***a***. Negative amplitudes are mirrored,

$$y = -Q(|x|) \qquad (35)$$

The generalisation is reflected by the amplitude dependent values $\lambda(x)$, $q(x)$, $r(x)$ in eq. (34). For a given set of representation levels $\ldots < r_{l-1} < r_l < r_{l+1} < \ldots$ and a given amplitude **x**, the pair of consecutive representation levels is selected that fulfils

$$r_{l-1} \leq x < r_l \qquad (36)$$

The value of the local representation level is then set to

$$r(x) = r_{l-1} \qquad (37)$$

The value of the local quantisation step-size results from

$$q(x) = q_l = r_l - r_{l-1} \qquad (38)$$

A straightforward extension of the rate-distortion concept detailed above yields for the local lambda parameter, very similar to eq. (20) ,

$$\lambda(x) = \lambda_l = 1 - \frac{\mu}{q_l^2} \cdot \log_2\left(\frac{P_{l-1}}{P_l}\right) \qquad (39)$$

(l = 1, ..., L)

Similar to eqs. (13), (14), the probabilities in eq. (39) depend on the lambda parameters,

$$P_0 = \int_0^{r_1 - \frac{\lambda_1}{2} \cdot q_1} p(x)dx \qquad (40)$$

and

$$P_l = \int_{r_l - \frac{\lambda_l}{2} \cdot q_l}^{r_{l-1} - \frac{\lambda_{l-1}}{2} \cdot q_{l-1}} p(x)dx \quad l \geq 1 \qquad (41)$$

Therefore, eq. (39) represents a system of non-linear equations for determining the lambda parameters $\lambda_1, \ldots, \lambda_L$. In general, this system can only be solved numerically.

However, eq. (39) can be simplified if the term $\log_2(P_{l-1}/P_l)$ is interpreted as the difference

$$I_l - I_{l-1} = \log_2\left(\frac{P_{l-1}}{P_l}\right) \qquad (42)$$

of optimum codeword lengths

$$I_l = -\log_2 P_l \qquad I_{l-1} = -\log_2 P_{l-1} \qquad (43)$$

associated with the representation levels r_l, r_{l-1}.

A practical implementation of the above will now be described.

Once the probability distribution, parametric or actual, of the unquantized coefficients is known, it is possible to choose a set of quantizer decision levels that will minimise the cost function **B**, because both the entropy **H** and the distortion **D** are known as functions of the decision levels for a given probability distribution. This minimization can be performed off-line and the calculated sets of decision levels stored for each of a set of probability distributions.

In general, it will be seen that the optimum value of λ corresponding to each decision level is different for different coefficient amplitudes. In practice,

it appears that the greatest variation in the optimum value of λ with amplitude is apparent between the innermost quantizer level (the one whose reconstruction level is 0) and all the other levels. This means that it may be sufficient in some cases to calculate, for each coefficient index and for each value (suitably quantized) of the probability distribution parameter, two values of λ, one for the innermost quantizer level and one for all the others.

A practical approach following the above description is shown in Figure 5.

The DCT coefficients are taken to a linear quantizer 52 providing the input to a histogram building unit 54. The histogram is thus based on linearly quantized versions of the input DCT coefficients. The level spacing of that linear quantizer 52 is not critical but should probably be about the same as the average value of **q**. The extent of the histogram function required depends on the complexity of the parametric representation of the pdf; in the case of a Laplacian or Gaussian distribution it may be sufficient to calculate the mean or variance of the coefficients, while in the 'zero excluded' Laplacian used in the Paper it is sufficient to calculate the mean and the proportion of zero values. This histogram, which may be built up over a picture period or longer, is used in block 56 as the basis of an estimate of the pdf parameter or parameters, providing one of the inputs to the calculation of λ in block 58.

Another input to the calculation of λ is from a set of comparators 60 which are in effect a coarse quantizer, determining in which range of values the coefficient to be quantized falls. In the most likely case described above, it is sufficient to compare the value with the innermost non-zero reconstruction level. The final input required to calculate λ is the quantizer scale.

In general, an analytical equation for λ cannot be obtained. Instead, a set of values can be calculated numerically for various combinations of pdf parameters, comparator outputs and quantizer scale values, and the results stored in a lookup table. Such a table need not be very large (it may, for example, contain fewer than 1000 values) because the optima are

not very sharp.

The value of λ calculated is then divided by 2 and added in adder 62 to the coefficient prior to the final truncation operation in block 64.

Instead of using variable codeword lengths that depend on the current probabilities according to eq. (43), a fixed table of variable codeword lengths $C_0, ..., C_L$ can be applied to simplify the process. The values of $C_0,..., C_L$ can be determined in advance by designing a single variable length code, ie. a Huffman code, for a set of training signals and bit rates. In principle, they can also be obtained directly from the MPEG2 variable-length code table. The only complication is the fact that MPEG2 variable-length coding is based on combinations of runs of zero coefficients terminated by non-zero coefficients.

One solution to this problem is to estimate 'equivalent codeword lengths' from the MPEG2 VLC tables. This can be done quite easily if one makes the assumption that the probability distributions of the DCT coefficients are independent of each other. Another possibility is to consider the recent past history of quantization within the current DCT block to estimate the likely effect of each of the two possible quantization levels on the overall coding cost.

Then, eq. (39) changes to

$$\lambda(x) = \lambda_l = 1 - \frac{\mu}{q_l^2}(C_l - C_{l-1}) \quad ,(l = 1, ..., L) \qquad (44)$$

The resulting distortion-rate optimised quantisation algorithm is essentially the same as detailed previously except that the lambda parameters are calculated either from eq. (39) or eq. (44) for each pair of horizontal and vertical frequency indices.

A simplified method of calculating $\lambda(\mathbf{x})$ will now be described, where only the local distortion is considered for each coefficient.

Here, we make use of the fact that the variable-length code (VLC) table used for a given picture in MPEG2 is fixed and known. This should simplify and make more accurate the calculations of the trade-off between bit rate and distortion. In particular, the calculations can be made on a

coefficient basis since the effect on the bit rate of the options for quantizing a particular coefficient is immediately known. The same is true (although a little more difficult to justify) of the effect on the quantizing distortion.

If we accept the assumptions implied in the above paragraph, then we can very simply calculate the value of the decision level to minimize the local contribution to the cost function **B**. This will in fact be the level at which the reduction in the bit count obtained by quantizing to the lower reconstruction level (rather than the higher level) is offset exactly by the corresponding increase in quantizing distortion.

If the two reconstruction levels being considered have indices **i** and **i + 1**, the corresponding codewords have lengths L_i, and L_{i+1}, and the quantizer scale is **q**, then:

(i) the reduction in bit count is $L_{i+1} - L_i$.

(ii) the local increase in distortion is $20 \log_{10} q(1 - \lambda/2) - 20 \log_{10} q\lambda/2$.

Combining these using the law linking distortion to bit rate, we have

$$6(L_{i+1} - L_i) = 20 \log_{10}(2/\lambda - 1) \qquad (45)$$

or, more simply

$$L_{i+1} - L_i = \log_2(2/\lambda - 1) \qquad (46)$$

leading to

$$\lambda = 2/(1 + 2^{(L_{i+1} - L_i)}) \qquad (47)$$

This elegant result shows that the value of λ depends here only on the difference in bit count between the higher and lower quantizer reconstruction levels.

The fact that the level of λ is now independent both of the coefficient probability distribution and the quantizer scale leads to the following, much

simplified implementation shown in Figure 6.

Here, the DCT coefficients are passed to the side-chain truncate block 70 before serving as the address in a coding cost lookup table 72. The value of lambda/2 is provided to adder 76 by block 74 and the output is truncated in truncate block 78.

There have been described a considerable number of ways in which the present invention may be employed to improve quantisation in a coder; still others will be evident to the skilled reader. It should be understood that the invention is also applicable to transcoding and switching.

The question will now be addressed of a two stage-quantiser. This problem is addressed in detail in the Paper which sets out the theory of so-called maximum a-posteriori (MAP) and the mean squared error (MSE) quantisers. By way of further exemplification there will now be described an implementation of the MAP and MSE quantiser for transcoding of MPEG2 [MPEG2] intra AC-coefficients that result from an 8x8 discrete cosine transform (dct).

The class of the first generation quantisers $y_1 = Q_1(x)$ specified by these equations is spanned by the quantisation step-size q_1 and the parameter λ_1; such a quantiser is called (q_1, λ_1)-type quantiser.

In the transcoder, the first generation coefficients y_1 are mapped onto the second generation coefficients $y_2 = Q_2(y_1)$ to further reduce the bit rate. Under the assumption of a (q_1, λ_1)-type quantiser in the first generation, eg. MPEG2 reference coder TM5 , it follows from the results set out in the Paper that the MAP quantiser $Q_{2,map}$ and the MSE quantiser $Q_{2,mse}$ can be implemented as a $(q_2, \lambda_{2,map})$-type and a $(q_2, \lambda_{2,mse})$-type quantiser, respectively. For both, the MAP and the MSE quantiser, the second generation step-size q_2 is calculated from the second generation parameters w_2 and $qscale_2$. However, there are different equations for calculating $\lambda_{2,map}$ and $\lambda_{2,mse}$.

With the results of the Paper, it follows that $\lambda_{2,map}$ can be calculated as

$$\lambda_{2,map} = \lambda_{2,ref} + (\mu_{map} - \lambda_1) \cdot \frac{q_1}{q_2} \qquad (48)$$

and $\lambda_{2,mse}$ as

$$\lambda_{2,mse} = 1 + (\mu_{mse} - \lambda_1) \cdot \frac{q_1}{q_2} \qquad (49)$$

The parameter $\lambda_{2,ref}$ can be changed in the range $0 \le \lambda_{2,ref} \le 1$ for adjusting the bit rate and the resulting signal-to-noise-ratio. This gives an additional degree for freedom for the MAP quantiser compared with the MSE quantiser. The value of $\lambda_{2,ref} = 0.9$ is particularly preferred. The parameter μ_{map} and the parameter μ_{mse} are calculated from the first generation quantisation step-size q_1 and a z-value,

$$\mu_{map} = -\frac{2}{\ln(z^{q_1})} \cdot \ln\left(\frac{2}{1 + z^{q_1}}\right) \qquad (50)$$

$$\mu_{mse} = -\frac{2}{\ln(z^{q_1})} \cdot \frac{1 - (1 - \ln(z^{q_1})) \cdot z^{q_1}}{1 - z^{q_1}} \qquad (51)$$

The amplitude range of the values that result from these equations can be limited to the range $0 \le \mu_{map}, \mu_{mse} \le 2$. Similarly, the amplitude range of the resulting values can be limited to $0 \le \lambda_{2,map}, \lambda_{2,mse} \le 2$.

The z-value has a normalised amplitude range, ie. $0 \le z \le 1$, and can be calculated either from the first generation dct-coefficients y_1 or from the original dct-coefficients x as described in the Paper. In the latter case, the z-value is transmitted as additional side information, eg. *user data*, along with

the first generation bit stream so that no additional calculation of **z** is required in the transcoder. Alternatively, a default **z**-value may be used. An individual **z**-value is assigned to each pair of horizontal and vertical frequency indices. This results in 63 different **z**-values for the AC-coefficients of an 8x8 dct. As a consequence of the frequency dependent **z**-values, the parameters $\lambda_{2,map}$ and $\lambda_{2,mse}$ are also frequency dependent, resulting in 63 $(q_2, \lambda_{2,map})$-type quantisers and 63 $(q_2, \lambda_{2,mse})$-type quantisers, respectively. Additionally, there are different parameter sets for the luminance and the chrominance components. The default **z**-values for the luminance and chrominance components are shown in Table 1 and Table 2 respectively.

TABLE 1

Normalised z-values, eg. 256 **x z**, for luminance (default)

	0	1	2	3	4	5	6	7
0		252	250	247	244	239	233	232
1	251	249	247	244	240	235	233	231
2	249	247	245	242	238	236	234	233
3	247	245	243	240	236	235	234	233
4	246	242	239	237	235	233	232	235
5	242	238	235	234	230	229	231	233
6	237	231	226	225	222	222	226	231
7	222	211	210	205	202	208	214	222

TABLE 2

Normalised **z**-values, ie. 256 **x z**, for chrominance (default)

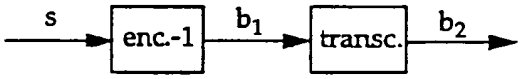
	0	1	2	3	4	5	6	7
0		248	242	230	212	176	158	179
1	246	240	233	219	193	154	156	177
2	239	233	224	209	180	148	154	173
3	229	221	211	196	163	141	150	166
4	219	208	198	181	153	133	143	166
5	207	193	182	171	140	126	143	161
6	193	176	162	154	127	118	137	163
7	169	145	148	129	102	108	127	158

For a description of preferred techniques for making available to subsequent coding and decoding processes, information relating to earlier coding and decoding processes, reference is directed to EP-A-0 765 576; EP-A-0 807 356 and WO-A-9803017.

Transcoding of MPEG-2 intra frames

1. Introduction

Transcoding is the key technique to further reduce the bit rate of a previously compressed image signal. In contrast to a standalone source encoder, a transcoder has only access to a previously compressed signal that already contains quantisation noise when compared to the original source signal. Thus, the bit stream output of the transcoder is the result of cascaded coding with a so called first generation encoder in the first stage followed by the transcoder, see Fig.1.

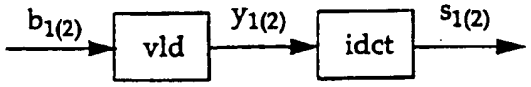


s = source signal, enc.-1 = first generation encoder, transc. = transcoder

b_1 , b_2 = first, second generation bit stream

Fig. 1: Cascaded coding as a result of first generation encoding and subsequent transcoding

It is assumed throughout that the first generation bit stream b_1 that defines the input of the transcoder and the second generation bit stream b_2 that represents the output are both MPEG-2 [MPEG-2] compliant. Hence, b_1 and b_2 can be passed on to a MPEG-2 decoder. In MPEG-2 motion compensating prediction (mcp) is combined with the discrete cosine transform (dct) in a hybrid coding algorithm [TM5-93]. In general both elements, mcp and dct coding, can be exploited for efficient transcoding. Compared to inter frames, i.e. P- and B-frames, mcp is switched off in intra frames (I-frames). Therefore, only dct coding can be exploited to transcode I-frames. In this paper we concentrate on MPEG-2 compatible transcoding of I-frames. A generalized block diagram of a MPEG-2 I-frame decoder is given in Fig. 2



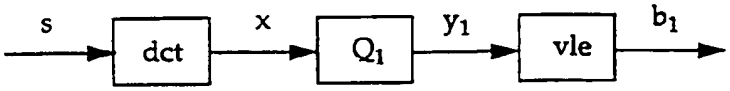
vld = variable length decoder, idct = inverse discrete cosine transform

$b_{1(2)}$ = first (second) generation bit stream, $y_{1(2)}$ = decoded dct-coeff. of first (second) generation

$s_{1(2)}$ = reconstructed image signals of first (second) generation

Fig.2 : Generalized MPEG-2 I-frame decoder

The corresponding first generation encoder and transcoder are detailed in Fig.3 and Fig.4, respectively.

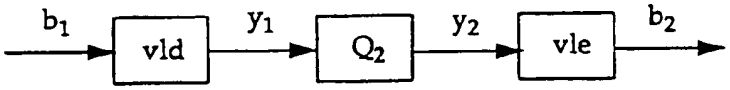


dct = discrete cosine transform, Q_1 = first generation quantiser, vle = variable length encoder

s = original source signal, x = original dct-coeff., y_1 = dct-coeff. of first generation

b_1 = first generation bit stream

Fig. 3 : Generalized MPEG-2 compatible first generation I-frame encoder



vld = variable length decoder, Q_2 = second generation quantiser, vle = variable length encoder

b_1 = first generation bit stream, y_1 = dct-coeff. of first generation

b_2 = second generation bit stream, y_2 = dct-coeff. of second generation

Fig. 4 : Generalized MPEG-2 compatible I-frame transcoder

As can be seen from Fig.4 the element of the transcoder to further reduce the bit rate is the second generation quantiser Q_2 . Hence, the fundamental issue of transcoding is the design of Q_2. To the author's knowledge, only a few publications have previously addressed this problem.

In [BT-94] and [Sarnoff-96] it is suggested to implement in Q_2 essentially the same quantiser characteristic that is used for Q_1 in the first generation encoder, Fig.3. In this case Q_1 and Q_2 have the same shape, e.g. a uniform quantiser characteristic, and differ only in their level of coarseness, i.e. the representation levels of Q_2 are more widely spaced compared to the representation levels of Q_1. Another approach to specify Q_2 is described in [Columbia-95]. Each DCT-coefficient y_1 of the first generation is checked, and a decision is made whether to retain or skip it, i.e. $y_2 = Q_2(y_1)$, where either $Q_2(y_1) = y_1$ or $Q_2(y_1) = 0$ holds. The decision whether to retain or skip is determined in a iterative optimisation procedure on a frame basis. Experimental results given in [Columbia-95] indicate that, to a large extent, high frequency dct-coefficents are skipped and low frequency dct-coefficients are retained. Therefore, this type of quantisation results in dct-based low pass filtering, which in general carries the risk of introducing visible block artefacts. Whilst this approach may be considered for transcoding for small bit rate changes between the first and the second generation, it appears questionable whether skipping of dct-coefficients functions well in general. Unfortunately, experimental results in [Columbia-95] are only given for transcoding between 4 and 3 Mbits/s, but not e.g. between 9 and 3 Mbit/s. In [Sarnoff-96] the authors show in their experimental results that, even for transcoding between 4 and 3 Mbit/s, skipping of dct-coefficients is inferior compared with the above re-quantisation where Q_1 and Q_2 have the same shape and differ only in their level of coarseness. However, different algorithms for skipping of dct-coefficients have been used in [Sarnoff-96] and in [Columbia-95].

This paper provides a theoretical analysis of the transcoding problem. The formal description of the second generation quantiser Q_2 includes the suggestions mentioned above of [BT-94][Columbia-95][Sarnoff-96] as special cases. From the results of the analysis, we derive theoretically the optimum quantiser characteristic Q_2 for both the mean-squared-error (mse) cost function and a so-called maximum-a-posteriori (map) cost function. The difference between these two cost functions is explained, along with pointing out in which case each cost function is more suitable. In order to efficiently apply the optimum mse- and map-quantiser characteristics in a transcoder, it is necessary to model the statistics of the original dct-coefficients x, see Fig.3. This paper proposes a parametric model. This model is first validated with real image data before it is used in the experiments to evaluate the mse- and map-quantiser characteristics. For reference, the results are compared with the performance of the quantiser characteristic of MPEG-2 test model TM5 [TM5-93].

The rest of the paper is organised as follows. In section 2, a formal desription of a MPEG-2 compatible quantiser is introduced. Two examples suitable for the quantiser Q_1 of a first generation encoder are discussed and compared. Section 3 focusses on the second generation quantiser Q_2

used in the transcoder. The problem of designing Q_2 is analysed by making a comparison to a reference quantiser $Q_{2,ref}$. The reference quantiser is used in a standalone encoder that bypasses the first generation stage and directly compresses the original signal to the desired bit rate of the second generation. In Section 4, two methods of designing Q_2 are proposed, explained, and compared. The first method minimises the mse cost function, the second minimises the map related costs. In Section 5 a parametric model to describe the statistics of the original dct-coefficients is introduced and validated with real image data. Based on this parametric model, an analytical evaluation for both the mse and the map cost function is carried out in section 6. Experimental results are discussed in section 7. Finally, conclusions and suggestions for future work are given in section 8.

2. MPEG-2 compatible quantisation in a first generation encoder

An MPEG-2 I-frame is partitioned into blocks of 8x8 samples of the original signal s. As shown in Fig.3, each block is submitted to the dct, resulting in 64 dct-coefficients x_i. The DC-coefficient x_0 is separately quantised and encoded. As the DC-quantiser characteristic is fixed for each frame, we concentrate on the AC-coefficients x_i, $i=1,2,...63$, with amplitude range $|x_i| \le 1024$. Without loss of generality, the frequency index i can be fixed for the following discussion. For clarity the frequency index i is therefore omitted. The original dct-coefficient x is passed to the first generation quantiser Q_1. The MPEG-2 standard [MPEG-2] specifies in its normative part the set of representation levels $y_1 = Q_1(x)$; in a MPEG-2 decoder, see Fig.2, a representation level y_1 is reconstructed as

$$y_1 = l_1 \cdot q_1 \quad , \tag{1}$$

with the quantisation step-size

$$q_1 = \frac{w_1 \cdot qscale_1}{16} . \tag{2}$$

The amplitude level l_1 can take an integer value out of the allowed amplitude range $|l_1| \le 2047$, and is transmitted as (8x8)-block data in the first generation bit stream b_1; b_1 includes as additional side information the values of w_1 and $qscale_1$ needed to calculate the quantisation step-size q_1. The value of w_1 can be set for each frame and depends on the frequency index, thus taking into account the frequency dependent properties of human visual perception. The value of $qscale_1$ does not depend on the frequency index and can be changed on a macroblock basis within each frame. A macroblock consists of four luminance and two co-sited chrominance blocks, each of 8x8 samples. A MPEG-2 compatible quantiser complies with the reconstruction rules of (1) and (2). Additionally, y_1 has to be rounded to an integer which is omitted in our discussion to ease the notation. For a given step-size q_1 there is no unique MPEG-2 compatible

quantiser $y_1 = Q_1(x)$ because of the remaining degree of freedom how to map the set of original samples x onto the given set of representation levels.

As an example, the quantiser characteristic can be specified for non-negative values x as

$$y_1 = Q_1(x) = \left\lfloor \frac{x}{q_1} + \frac{\lambda_1}{2} \right\rfloor \cdot q_1, \tag{3}$$

where the floor function $\lfloor a \rfloor$ extracts the integer part of the given argument a. The quantiser characteristic of (3) can be mirrored for negative values of x,

$$y_1 = -Q_1(|x|) \tag{4}$$

The parameter λ_1 in eq. (3) determines how the positive x-axis is partitioned into half-open intervals $[d_{1l}, d_{1(l+1)})$. Every interval is defined by two consecutive decision levels d_{1l} with $d_{10} = 0$ for $l = 0$ and

$$d_{1l} = \left(l - \frac{\lambda_1}{2}\right) \cdot q_1 \tag{5}$$

for $l \geq 1$. According to (3), each $x \in [d_{1l}, d_{1(l+1)})$ is mapped onto the same representation level r_{1l}

$$r_{1l} = l \cdot q_1 \tag{6}$$

see Fig. 5.

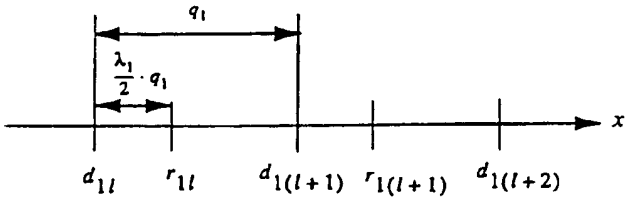


Fig. 5: Decision and representation levels of the quantiser in eq. (3)

The value of λ_1 can be tuned to the cost function that is used to measure the quantiser performance. Here, two types of performance are considered.

(i) mse performance

The mean-squared-error (mse) is a familiar cost function to measure the quantiser performance. In this case the value of λ_1 is determined by minimising the expectation value

$$E\left[(x - y_1)^2\right]. \tag{7}$$

In order to minimise (7), each orginal dct-coefficient x has to be mapped onto the nearest representation level $y_l = r_{1l}$. Thus, without applying the calculus for differentiation one can conclude: the corresponding decision levels d_{1l} that minimise (7) are defined by the arithmetic mean values

$$d_{1l} = \frac{r_{1(l-1)} + r_{1l}}{2} \tag{8}$$

for the given set of representation levels r_{1l}. The result in eq. (8) can be regarded as the first half of the celebrated Lloyd-Max quantiser design rule described in [Lloyd-57] and [Max-60] which in its second half requires each representation level r_{1l} to coincide with the local centroid of the corresponding bin $[d_{1l}, d_{1(l+1)})$. However, the latter can in general not be achieved for the signal-independent representation levels of eq. (6) because the local centroids depend on the probability distribution of the original dct-coefficients x, and are therefore signal-dependent. With eqs. (6) and (8) it follows from eq. (5) that $\lambda_1 = 1$ is the solution for the mse cost function.

(ii) rate-distortion performance

In a rate-distortion sense it is more suitable to minimise the mse term of eq. (7) subject to a given bit rate constraint for the first generation bit stream b_1. In this case the bit rate needed to encode the first generation dct-coefficients y_1 is not allowed to exceed a preset value H. Let $H_1(\lambda_1)$ denote the resulting bit rate for an adjustment of the decision levels d_{1l} according to eq. (5). It then follows from the method of Lagrange multipliers [Heuser-82] that the optimum value of λ_1 can be found by minimising the extended cost function

$$E\left[(x - y_1)^2\right] + \mu \cdot H_1(\lambda_1). \tag{9}$$

The Lagrange multiplier μ in eq. (9) is determined by the constraint

$$H_1(\lambda_1) \leq H. \tag{10}$$

There are two extreme cases. For $H \to \infty$ the Lagrange multiplier aproaches zero, $\mu \to 0$, and eq. (9) coincides with eq. (7), i.e. no attention is paid to the resulting bit rate and only the mse is minimised. For $H \to 0$ no dct-coeffcients y_1 can be coded. As a consequence, the decision level d_{1l} already approaches infinity for $l = 1$, i.e. $d_{11} \to \infty$ so that all original coefficients x are quantised to zero. If the latter case with $d_{11} \to \infty$ is only applied to particular blocks or to a selected range of frequency indices, e.g. all high frequency dct-coefficients, then, the optimisation of eqs. (9) and (10) results in skipping of dct-coefficients.

The first twenty-five frames of the CCIR 601 test signal 'mobile' have been coded to evaluate both the mse and the rate-distortion performance of the quantiser specified in eqs. (3) and (4). Two values of λ_1 have been used: a.) $\lambda_1 = 1$ and b.) $\lambda_1 = 0.75$. The weighting matrix proposed in MPEG-2 test model TM5 [TM5-93] has been applied for a frequency dependent quantisation. Fig. 6 shows the resulting peak-signal-to-noise-ratio (PSNR) as a function of $qscale_1$, see eq. (2) for the meaning of $qscale_1$.

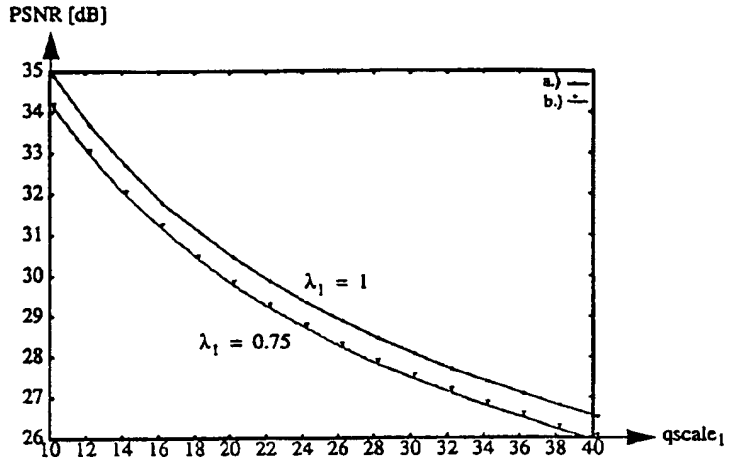


Fig. 6: PSNR/mse performance of the quantiser of eqs. (3) and (4) for the test signal 'mobile'

In accordance with the theoretical results the largest PSNR values are achieved for $\lambda_1 = 1$. The PSNR values drop by about 0.4 dB if the value is changed to $\lambda_1 = 0.75$. However, $\lambda_1 = 0.75$ is the better choice in terms of rate-distortion performance for small and medium bit rates as is revealed in Fig. 7. For a given distortion of e.g. PSNR = 30 dB the resulting bit amount is approx. 450000 bit/frame in case of $\lambda_1 = 1$ and approx. 410000 bit/frame in case of $\lambda_1 = 0.75$, this results in a bit saving of approx. 9% in the latter case. Conversely, for a given bit amount of 450000 bit/frame the PSNR value can be increased by about 0.5 dB from 30 dB to 30.5 dB when changing from $\lambda_1 = 1$ to $\lambda_1 = 0.75$. As explained above, when the bit rate is infinite, again the rate-distortion performance of $\lambda_1 = 1$ is best. Hence, one can expect an intersection between the rate-distortion curves of Fig. 7 when the bit rate is increased. This is shown in Fig. 8. For bit rates larger than approx. 1 Megabit/frame which corresponds to PSNR values larger than 38 dB the results are in favour of $\lambda_1 = 1$. However, the difference to the result of $\lambda_1 = 0.75$ is rather small.

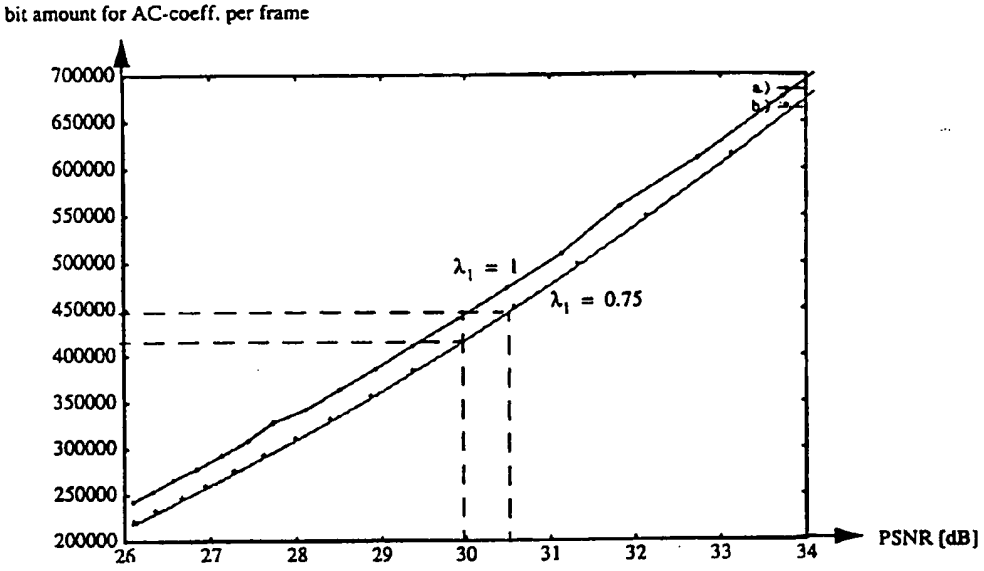


Fig. 7: Rate-distortion performance in case of small and medium bit rates for the quantiser of eqs. (3) and (4), test signal 'mobile'

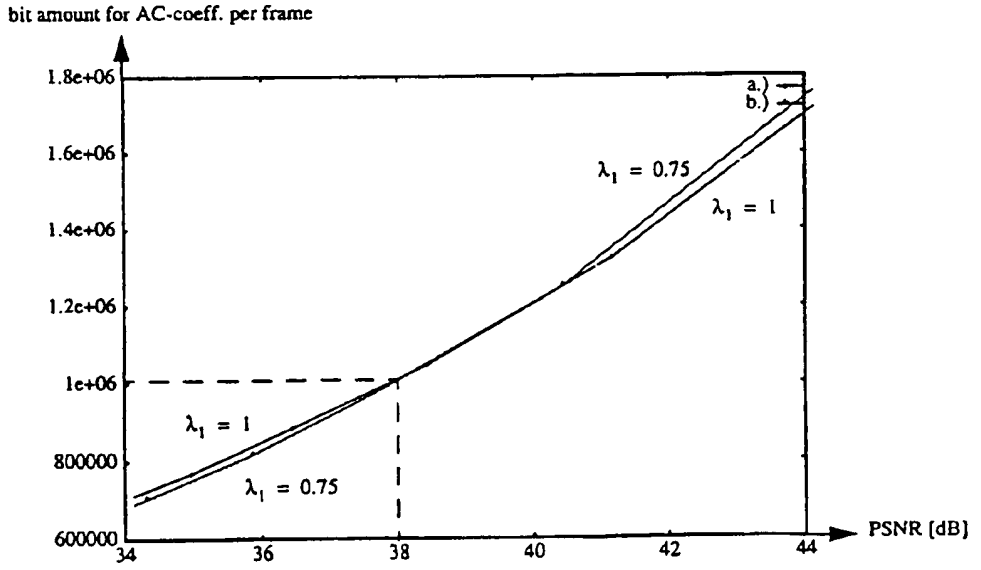


Fig. 8: Rate-distortion performance in case of high bit rates for the quantiser of eqs. (3) and (4), test signal 'mobile'

The value $\lambda_1 = 0.75$ is the 'intended' value for the quantiser of the MPEG-2 test model TM5. Due to a simplified calculation of eqs. (3) and (4) involving integer rounding operations, the value of λ_1 that is used in TM5 depends on the value of $qscale_1$, for further details the reader is refered to the TM5 description in [TM5-93]. The functional relationship between λ_1 and $qscale_1$ of TM5 is shown in Fig. 9.

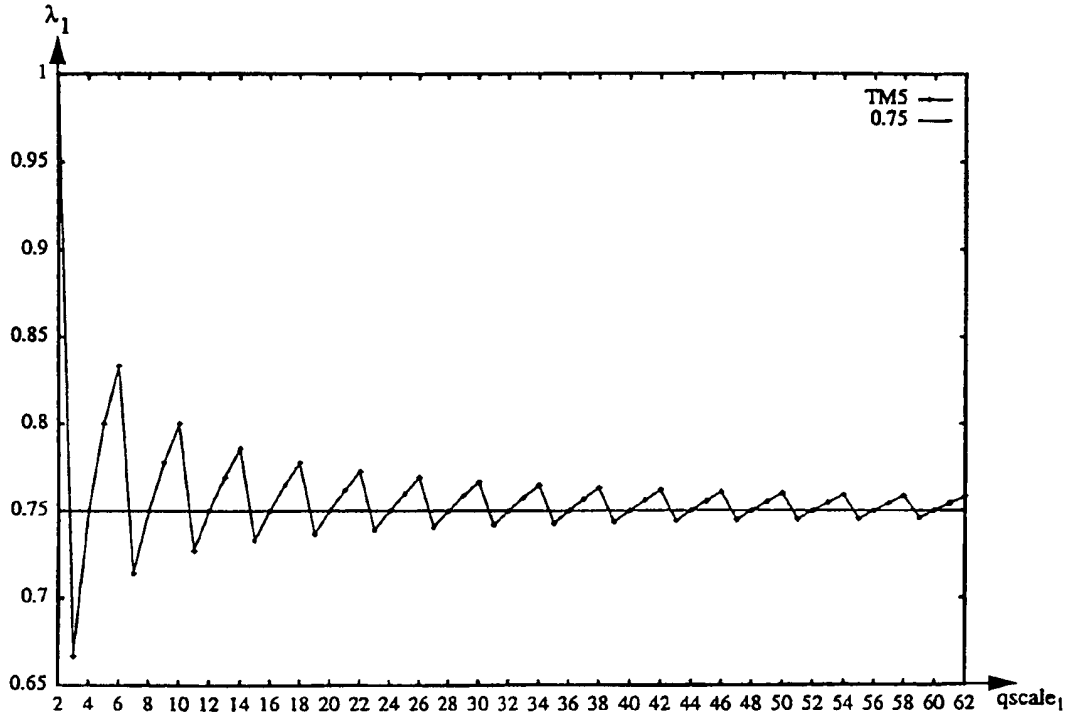


Fig. 9: MPEG-2 test model TM5, functional relationship between λ_1 and $qscale_1$

For $qscale_1 = 2$ the resulting lambda value is $\lambda_1 = 1$, this is fine because a small $qscale_1$ value corresponds to a high bit rate, and then $\lambda_1 = 1$ is the best choice. With increased $qscale_1$ the λ_1 values approach the 'intended' value 0.75. This value is exactly matched, i.e. $\lambda_1 = 0.75$, if $qscale_1$ is a multiple of 4, e.g. $qscale_1 = 4,8,12,16,...$

Of course, the experimental results give no evidence that $\lambda_1 = 0.75$ is the optimum value in the rate-distortion sense of eqs. (9) and (10). The tuning of λ_1, or more generally, the adjustment of the decision levels for a given set of representation levels to improve the rate-distortion performance of the first generation quantiser is another issue and beyond this paper's scope. In the following sections we concentrate on the second generation quantiser and investigate how the degree of freedom that lies in the adjustment of the decision levels can be exploited

for efficient transcoding.

3. MPEG-2 compatible quantisation in a transcoder - the transcoding problem

In contrast to the first generation quantiser Q_1 that has access to the original dct-coefficients x, the second generation quantiser Q_2 used in the transcoder has only access to the first generation dct-coefficients y_1. Thus, Q_2 maps the first generation dct-coefficients onto the second generation,

$$y_2 = Q_2(y_1). \tag{11}$$

With

$$y_1 = Q_1(x) \tag{12}$$

for the first generation quantiser, the relationship between the original dct-coefficients and the second generation becomes

$$y_2 = Q_2(Q_1(x)). \tag{13}$$

Ideally, the result of cascaded quantisation in eq. (13) should be identical to the output of a reference quantiser $Q_{2,ref}$ that has access to the original dct-coefficients,

$$y_{2,\,ref} = Q_{2,\,ref}(x). \tag{14}$$

The reference quantiser is used in a standalone encoder that by-passes the first generation stage and directly compresses the original signal to the desired bit rate of the second generation. For $y_2 = y_{2,\,ref}$ there is no additional loss due to transcoding.

Therefore, we start analysing the transcoding problem with the investigation for which cases $y_2 = y_{2,\,ref}$ is achievable. As an example, Fig. 10 shows the quantiser characteristics of the first generation and of the reference for transcoding. In this example, the two basic cases of transcoding occur.

Case 1: In the first case the half-open interval $[d_{1l}, d_{1(l+1)})$ on the x-axis is considered. Each $x \in [d_{1l}, d_{1(l+1)})$ is mapped by the first generation quantiser onto the representation level $y_1 = Q_1(x) = r_{1l}$. As a consequence, no matter on which second generation value y_2 this representation level is mapped by the transcoder's characteristic $y_2 = Q_2(y_1 = r_{1l})$, the resulting characteristic of eq. (13) will always be constant over the entire interval $[d_{1l}, d_{1(l+1)})$. However, the reference quantiser of eq. (14) changes the representation level over the interval $[d_{1l}, d_{1(l+1)})$. Each $x \in [d_{1l}, d_{L,\,ref})$ is mapped onto the representation level $y_{2,\,ref} = Q_{2,\,ref}(x) = r_{2(L-1)}$ while each $x \in [d_{L,\,ref}, d_{1(l+1)})$ is mapped onto r_{2L}. The dilemma for the transcoder is that only the entire interval $[d_{1l}, d_{1(l+1)})$ can be mapped either on $r_{2(L-1)}$ or on r_{2L}. Clearly, no matter how the final choice is done, $y_2 = y_{2,\,ref}$ is not achievable for all $x \in [d_{1l}, d_{1(l+1)})$.

Case 2: In the second case the interval $[d_{1(l+1)}, d_{1(l+2)})$ on the x-axis is considered. In this case the reference quantiser maps the entire interval $[d_{1(l+1)}, d_{1(l+2)})$ on the representation level r_{2L}. Hence, $y_2 = y_{2,\,ref}$ is achievable for all $x \in [d_{1(l+1)}, d_{1(l+2)})$ if the transcoder applies the mapping

$$r_{2L} = y_2 = Q_2(y_1 = r_{1(l+1)}) \,. \tag{15}$$

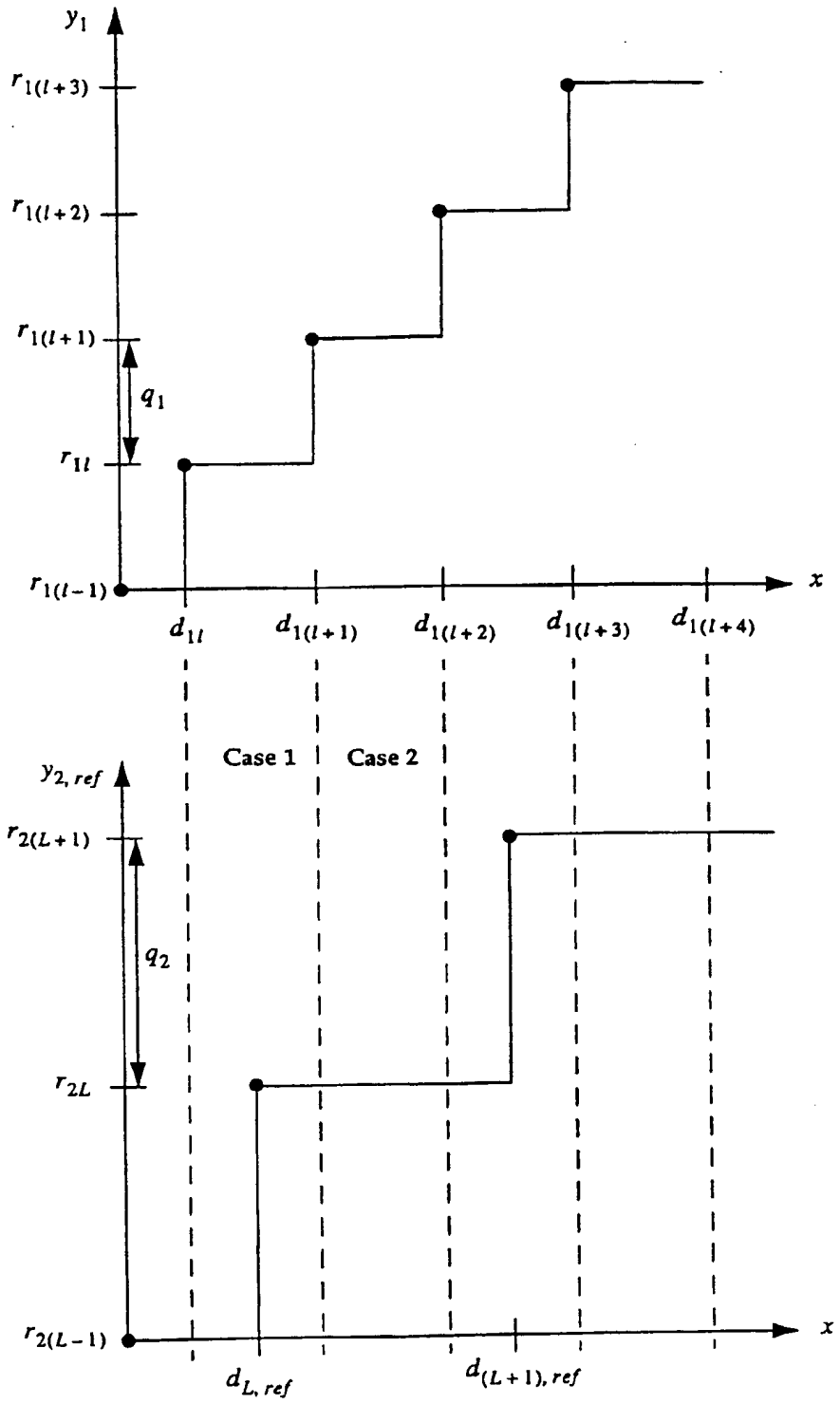


Fig. 10: Quantiser characteristics of first generation and of the reference for transcoding

Case 1 must be avoided in order to fulfil $y_2 = y_{2,\,ref}$ for the whole x-axis. This is accomplished if the set of decision levels $\{d_{L,\,ref}\}$ of the reference quantiser forms a subset of the set of decision levels $\{d_{1l}\}$ of the first generation quantiser,

$$\{d_{L,\,ref}\} \subseteq \{d_{1l}\}. \tag{16}$$

For the parametric description of the decision levels in eq. (5), condition (16) can be translated into an equivalent condition that involves the quantisation step-sizes q_1 and q_2 of the first and the second generation, respectively,

$$\frac{q_2}{q_1} = \frac{\left(n_0 + 1 - \frac{\lambda_1}{2}\right)}{1 - \frac{\lambda_{2,\,ref}}{2}} = k \in \{1, 2, 3, \ldots\}. \tag{17}$$

Eq. (17) reads as follows. For given parameters λ_1 and $\lambda_{2,\,ref}$ of the first generation and the reference quantiser, respectively, $y_2 = y_{2,\,ref}$ is achievable for the whole x-axis if there exists a positive integer n_0 such that the middle term of (17) results in another positive integer k which at the same time has to be equal to the ratio $\frac{q_2}{q_1}$ of the step-sizes.

As an example, in the case of the mse cost function for the first generation and the reference quantiser with $\lambda_1 = \lambda_{2,\,ref} = 1$, eq. (17) becomes $\frac{q_2}{q_1} = (2 \cdot n_0 + 1) = k$. This means that the ratio $\frac{q_2}{q_1}$ has to be equal to an odd valued integer, e.g. 3, 5, 7, ...

In general for arbitrary q_1 and q_2, condition (17) cannot be met, and as a consequence, there is an additional loss due to transcoding. This loss can be described by Case 1. It can be derived from Fig. 10 that the mismatch between y_2 and $y_{2,\,ref}$ is especially large if the decision level $d_{L,\,ref}$ coincides with the centre of the interval $[d_{1l}, d_{1(l+1)})$. For the above example of $\lambda_1 = \lambda_{2,\,ref} = 1$ this is indicated by an even ratio $q_2/q_1 = 2, 4, 6, \ldots$

However, Case 1 does not occur throughout if eq. (17) is not fulfilled. The decision intervals of the first generation quantiser can be partitioned into two classes, those who belong to Case 1, e.g. $[d_{1l}, d_{1(l+1)})$ in Fig. 10, and those who belong to Case 2, e.g. $[d_{1(l+1)}, d_{1(l+2)})$ in Fig. 10. In general, the resulting two classes do not have equal number of assigned intervals. The percentage of intervals that belong to Case 1 depends on q_1 and q_2 as well as on λ_1 and $\lambda_{2,\,ref}$. The percentage of Case 1 intervals decreases asymptotically to zero as the ratio $\frac{q_2}{q_1}$ tends to infinity, compare Fig. 10. Indeed, $\frac{q_2}{q_1} = \infty$ can be regarded as a special case of eq. (17) for $n_0, k \to \infty$, thus, Case 2 intervals are guaranteed throughout. For $q_1 > 0$ this results in $q_2 = \infty$ which can be interpreted as skipping of dct-coefficients in the transcoder and as well

in the reference quantiser. However, skipping of dct-coefficients may not be the default for a sophisticated reference quantiser to achieve a good rate-distortion performance, e.g. a reference quantiser defined by eqs. (3) and (4). Therefore, techniques are required that minimise the additional loss due to transcoding in Case 1 intervals. This problem will be addressed in the following section.

4. Design of the second generation quantiser Q_2 used in the transcoder

From the results of the previous section it follows that the mapping of eq. (15) should be applied in the transcoder for Case 2 intervals. Therefore, this section concentrates on Case 1 intervals. Resuming the discussion of Case 1, the transcoder has to take a binary decision for the interval $[d_{1l}, d_{1(l+1)})$ in Fig. 10. As the interval $[d_{1l}, d_{1(l+1)})$ is represented by the first generation level $y_1 = Q_1(x) = r_{1l}$ in the transcoder, the decision is to map $y_1 = r_{1l}$ either onto the second generation representation level $y_2 = r_{2(L-1)}$ or onto $y_2 = r_{2L}$. This decision can be taken based on the minimisation of a cost function. Here, two cost functions are considered.

4.1 MSE cost function

Similar to eq. (7) in section 3, the mean-squared-error (mse) cost function can be applied, resulting in the expectation value

$$E\left[(x-y_2)^2\right]. \tag{18}$$

Thus, the objective of this cost function is to minimise the mse between the original dct-coefficients x and the second generation coefficients y_2 that follow from the transcoder characteristic $y_2 = Q_2(y_1)$. For the mse cost function of eq. (18), the corresponding reference quantiser $Q_{2,\,ref}$ is defined by eqs. (3) and (4) with the parameter $\lambda_{2,\,ref} = 1$ and the second generation step-size q_2. Hence, the decision level $d_{L,\,ref}$ in Fig. 10 is related to the representation levels by eq. (8), i.e.

$$d_{L,\,ref} = \frac{r_{2(L-1)} + r_{2L}}{2}. \tag{19}$$

As the transcoder has access to the first generation coefficients y_1, this information can be exploited to minimise the term of eq. (18). Therefore, the mse cost function is expanded by introducing a conditional expectation value that depends on y_1,

$$E\left[(x-y_2)^2\right] = E\left[E\left[(x-y_2)^2 \middle| y_1\right]\right]. \tag{20}$$

With eq. (20) the resulting binary decision rule can be written for the Case 1 interval $[d_{1l}, d_{1(l+1)})$ in Fig. 10 as follows,

$$y_2 = Q_2(y_1 = r_{1l}) = \begin{cases} r_{2(L-1)} \\ r_{2L} \end{cases} \quad \text{if} \quad \begin{matrix} E\left[(x - r_{2(L-1)})^2 \middle| y_1\right] < E\left[(x - r_{2L})^2 \middle| y_1\right] \\ E\left[(x - r_{2(L-1)})^2 \middle| y_1\right] > E\left[(x - r_{2L})^2 \middle| y_1\right] \end{matrix} . \tag{21}$$

The conditional expectation values in eq. (21) depend on the probability density function (pdf) $p(x)$ of the original dct-coefficients x. The pdf determines the probability P_{1l} that a single value x falls in the interval $[d_{1l}, d_{1(l+1)})$, and is mapped onto the first generation level $y_1 = Q_1(x) = r_{1l}$,

$$P_{1l} = \int_{d_{1l}}^{d_{1(l+1)}} p(x)dx. \tag{22}$$

With the definition of a local centroid,

$$c_{1l} = \frac{\int_{d_{1l}}^{d_{1(l+1)}} (x \cdot p(x))dx}{P_{1l}}, \tag{23}$$

and the decision level of eq. (19), the decision rule of eq. (21) can be re-stated after a straightforward calculation as

$$y_2 = Q_2(y_1 = r_{1l}) = \begin{cases} r_{2(L-1)} \\ r_{2L} \end{cases} \quad \text{if} \quad \begin{matrix} c_{1l} < d_{L,\,ref} \\ c_{1l} > d_{L,\,ref} \end{matrix} . \tag{24}$$

In the case of $c_{1l} = d_{L,\,ref}$ in eq. (24), an arbitrary decision between $r_{2(L-1)}$ and r_{2L} can be made. However, the decision might then be in favour of the lower level, i.e. $r_{2(L-1)}$, because in general lower amplitude levels correspond to MPEG-2 codewords of smaller length.

4.1.1 Implementation of the mse cost function

There are several ways to implement (24) in a transcoder. Here, we outline a straightforward implementation of the mse cost function as a quantiser characteristic $y_2 = Q_2(y_1)$. Given the first generation coefficient $y_1 = r_{1l}$, the transcoder can compute the corresponding interval $[d_{1l}, d_{1(l+1)})$ on the x-axis, e.g. for the first generation quantiser of eqs. (3), (4) one obtains

$$d_{1l} = r_{1l} - \left(\frac{\lambda_1}{2} \cdot q_1\right), \tag{25}$$

and

$$d_{1(l+1)} = d_{1l} + q_1 . \tag{26}$$

In passing we note that the first generation quantiser step-size q_1 required in eqs. (25) and (26) has to be transmitted as side information in any MPEG-2 bit stream. However, the parameter λ_1 is not specified in MPEG-2, and must therefore be additionally signalled, e.g. as *user data* in the first generation bit stream, see [MPEG-2] for the definition of *user data*.

Having computed the interval $[d_{1l}, d_{1(l+1)})$, the corresponding second generation representation levels can be determined with the reference quantiser $Q_{2,\,ref}$, see also Fig. 10,

$$r_{2(L-1)} = Q_{2,\,ref}(x = d_{1l}) = \left\lfloor \frac{d_{1l}}{q_2} + \frac{\lambda_{2,\,ref}}{2} \right\rfloor \cdot q_2 , \tag{27}$$

and

$$r_{2L} = Q_{2,\,ref}(x = d_{1(l+1)}) = \left\lfloor \frac{d_{1(l+1)}}{q_2} + \frac{\lambda_{2,\,ref}}{2} \right\rfloor \cdot q_2 , \tag{28}$$

with $\lambda_{2,\,ref} = 1$ for the mse cost function. With eqs. (27) and (28), the corresponding decision level $d_{L,\,ref}$ can be calculated from eq. (19).

The pdf $p(x)$ is needed to calculate the local centroid c_{1l} according to eqs. (22) and (23). In order to save the amount of additional side information, the transcoder can apply a model description for $p(x)$, involving only a few parameters. A parametric model suitable to describe the statistics of the original dct-coefficients x will be detailed and validated in the next section 5.

Now, all parameters are available, i.e. $r_{2(L-1)}$, r_{2L}, $d_{L,\,ref}$ and c_{1l}, to apply the decision rule of eq. (24). A more compact form of (24) in the sense of a quantiser characteristic can be

stated as follows,

$$y_2 = Q_{2,\,ref}(x = c_{1l}) = \left\lfloor \frac{c_{1l}}{q_2} + \frac{1}{2} \right\rfloor \cdot q_2 . \tag{29}$$

Eq. (29) specifies the second generation coefficient y_2 as a function of the local centroid c_{1l}. However, as we are interested rather in the transcoder's quantiser characteristic $y_2 = Q_2(y_1)$, the local centroid has to be related to the first generation coefficient y_1. As in eq. (25), the local centroid c_{1l} can be specified by a parameter μ_{1l} that relates the centroid to the decision level d_{1l},

$$d_{1l} = c_{1l} - \left(\frac{\mu_{1l}}{2} \cdot q_1\right). \tag{30}$$

Clearly, for $\mu_{1l} = 1$ the local centroid coincides with the centre of the decision interval $[d_{1l}, d_{1(l+1)})$ of length q_1. From eqs. (25) and (30) one deduces

$$c_{1l} = (y_1 = r_{1l}) + \left(\frac{\mu_{1l} - \lambda_1}{2} \cdot q_1\right). \tag{31}$$

After inserting (31) in (29) and by defining of

$$\lambda_2 = 1 + (\mu_{1l} - \lambda_1) \cdot \frac{q_1}{q_2}, \tag{32}$$

one obtains the desired result,

$$y_2 = Q_2(y_1 = r_{1l}) = \left\lfloor \frac{y_1}{q_2} + \frac{\lambda_2}{2} \right\rfloor \cdot q_2 . \tag{33}$$

Thus, the transcoder's quantiser characteristic of eq. (33) is essentially the same as the one of the first generation of eq. (3). However, as can be seen from (32), the parameter λ_2 is in general not a constant and depends on the actual value of μ_{1l} which can change with varying y_1. Thus, in contrast to the first generation parameter λ_1, the second generation parameter λ_2 depends in general on the actual input value y_1 of the transcoder's quantiser.

Nevertheless, in special cases depending on the pdf $p(x)$, the parameter λ_2 can become a constant, e.g. for a uniform pdf, i.e. $p(x) = const$, it follows from eqs. (22), (23) and (30) that $\mu_{1l} = 1$ throughout, and as a consequence of eq. (32) λ_2 becomes also a constant. Another special case with constant parameter λ_2 occurs if $\mu_{1l} = \lambda_1$ holds in eq. (32). In this case each first generation coefficient y_1 coincides with the corresponding local centroid c_{1l}, see eq. (31), resulting in $\lambda_2 = 1$ for the second generation.

4.2 From the mse to the map cost function

In addition to eq. (33), there is another instructive interpretation of the mse decision rule. We therefore have to re-state eq. (24) by substituting the local centroid c_{1l} with the parameter μ_{1l} according to eq. (30). After re-ordering one obtains

$$y_2 = Q_2(y_1 = r_{1l}) = \begin{cases} r_{2(L-1)} & \text{if } \frac{d_{L,ref} - d_{1l}}{q_1} > \frac{\mu_{1l}}{2} \\ r_{2L} & \text{if } \frac{d_{L,ref} - d_{1l}}{q_1} < \frac{\mu_{1l}}{2} \end{cases}. \tag{34}$$

With Fig. 10, the term $\frac{d_{L,ref} - d_{1l}}{q_1}$ in eq. (34) can be interpreted as the a-posteriori probability P_{uni} that the original dct-coefficient x falls in the interval $[d_{1l}, d_{L,ref})$ given the first generation coefficient $y_1 = r_{1l}$ in the case of a uniform pdf $p(x) = const$ over the interval $[d_{1l}, d_{1(l+1)})$ of length q_1, i.e.

$$P_{uni} = P[x \in [d_{1l}, d_{L,ref}) \,|\, y_1 = r_{1l}, p(x) = const] = \frac{d_{L,ref} - d_{1l}}{q_1}. \tag{35}$$

This is a curious result. Although the pdf $p(x)$ is non-uniform in general, eq. (34) suggests one computes P_{uni} and compares this value with the threshold $\frac{\mu_{1l}}{2}$.

In general, the a-posteriori probability P_{map} depends on the pdf as follows,

$$P_{map} = P[x \in [d_{1l}, d_{L,ref}) \,|\, y_1 = r_{1l}] = \frac{\int_{d_{1l}}^{d_{L,ref}} p(x)dx}{P_{1l}}, \tag{36}$$

where the denominator P_{1l} is defined as in eq. (22). The complementary a-posteriori probability is given by

$$\overline{P_{map}} = P[x \in [d_{L,ref}, d_{1(l+1)}) \,|\, y_1 = r_{1l}] = 1 - P_{map}. \tag{37}$$

For the special case of $p(x) = const$ the a-posteriori probability P_{map} of eq. (36) becomes identical to P_{uni} of eq. (35). Also, the parameters that are related to the local centroids then become $\mu_{1l} = 1$, throughout. Hence for this special case, the decision rule of eq. (34) can be restated with eqs. (36) and (37) as

$$y_2 = Q_2(y_1 = r_{1l}) = \begin{cases} r_{2(L-1)} & \text{if } P_{map} > \overline{P_{map}} \\ r_{2L} & P_{map} < \overline{P_{map}} \end{cases}. \quad (38)$$

In general for a non-uniform pdf, i.e. $p(x) \neq const$, the decision rule of eq. (38) does not coincide with the mse decision rule of eq. (34). The decision in (38) is based on the maximum a-posteriori (map) probability which is either P_{map} or $\overline{P_{map}}$. Therefore, eq. (38) is called the map decision rule.

Another important difference between the mse decision rule of eq. (34) and the map decision rule of eq. (38) is that the reference quantiser $Q_{2,\,ref}$ is fixed in the mse case but can be arbitrary in the map case. Note that $Q_{2,\,ref}$ determines the second generation decision level $d_{L,\,ref}$ in eqs. (36) and (37), and has therefore a significant impact on the a-posteriori probabilities. As an example, the map decision rule can be applied to the class of reference quantisers $Q_{2,\,ref}$ of eqs. (3) and (4) which is spanned by the parameter $\lambda_{2,\,ref}$. In the special case of $\lambda_{2,\,ref} = 1$, one obtains the reference quantiser of the mse cost function; however as explained earlier, the map decison rule will then only coincide with the mse decison rule if, additionally, the pdf is constant.

The degree of freedom to select the reference quantiser for the map decision rule can be exploited by specifying a characteristic $Q_{2,\,ref}$ that results in a sophisticated rate-distortion performance. Thus, the map decision rule can take into account not only the mse but also the resulting bit amount which is more suitable in a rate-distortion sense compared with the mse decision rule.

As the mse decision rule of eq. (34) minimises the cost function of eq. (18), one may now consider what cost function is minimised by the map decision rule of eq. (38). It is not difficult to verify that the map decision rule minimises the cost function

$$E\left[(y_{2,\,ref} - y_2)^2\right], \quad (39)$$

where y_2 and $y_{2,\,ref}$ are defined as in eqs. (13) and (14), respectively. Interestingly, it is again a mean-squared-error that we end up with. However, the mse of eq. (39) is more suitable in a rate-distortion sense than the mse of eq. (18).

Finally in this section, we note that both the mse and the map cost function belong to the family of Bayesian cost functions [Melsa-Cohn].

5. A parametric model to describe the statistics of the dct-coefficients

The transcoder needs to know the pdf $p(x)$ of the original dct-coefficients x in order to apply both the mse and the map cost functions. Basically, there are two possible approaches. In the first case, $p(x)$ is transmitted as additional side information along with the first generation bit stream. Second, no additional side information is transmitted, and $p(x)$ is estimated from the first generation dct-coefficients y_1. A parametric model is required that involves only a few parameters to limit the amount of additional side information in the first case. Also in the second case a parametric model with only a few parameters is needed because the reconstructed coefficients y_1 may not cover a sufficient amplitude range and/or their number may not be sufficient to achieve a reliable estimate for many parameters; this problem is known as 'context dilution' [Rissanen et. al. - 96]. In the following steps, a common parametric model will be derived that is suitable for both cases.

In the first step, we propose to model $p(x)$ as a Laplacian-like pdf

$$p_L(x) = \frac{\alpha}{2} \cdot e^{-\alpha|x|}, \tag{40}$$

which can be described by a single positive parameter α. A priori, the AC-coefficients of an 8x8 block cannot be assumed to share the same distribution. Therefore, an individual α value is specified for each AC-frequency index, resulting in 63 parameters in total.

Due to the discrete nature of the dct-coefficients, the probability P_{1l} to encounter a representation level $y_1 = r_{1l} = l \cdot q_1$ is given by eq. (22). For the special case of $q_1 = 1$, P_{1l} can be considered the probability of the original dct-coeffcients x. A histogram for ten consecutive frames of the CCIR 601 test signal 'mobile' has been computed in order to check whether the parametric model of (40) is suitable for the original dct-coefficients x. Due to the symmetry assumption for positive and negative amplitudes in eq. (40), the relative frequencies $\frac{n_l}{N}$ of absolute amplitude levels $|x| = l$ have been measured, where $0 \le l \le 1024$ and $N = 10 \times 6480$ for each AC frequency index. The results are similar for all 63 AC frequencies, three examples for selected horiziontal and vertical frequency indices are shown in Fig. 11. Each curve has its maximum at $|x| = l = 1$ and decays rapidly with increasing l. The maximum at $|x| = l = 1$ is larger and sharper peaked for higher frequencies, indicating a decreasing variance for increasing frequencies. Similar results have been obtained for other test signals. Qualitatively, the type of curve shown in Fig. 11 can be generated with eq.(40) by appropriate adjustment of the parameter α.

In order to get the best parameter fit one can take advantage of the fact that the transcoder needs to know the pdf $p(x)$ only for x values that are mapped onto non-zero representation levels $y_1 \neq 0$ by the first generation quantiser. This is because the first generation coefficients $y_1 = 0$ are always mapped onto the second generation coefficients $y_2 = 0$. Hence, the pdf has to be modelled only for $|x| \geq d_{11}$, where d_{11} is the smallest positive first generation decision level.

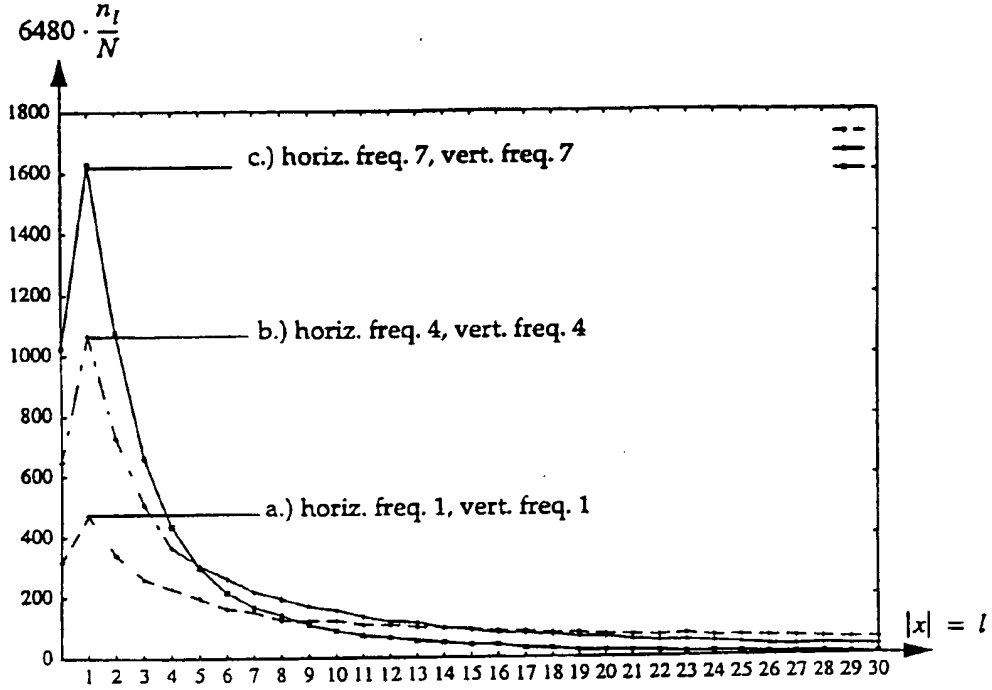


Fig. 11: Histogram for absolute amplitude levels of original dct-coefficients for different horizontal and vertical frequency indices, test signal 'mobile'

For the parametric description of eq. (5), one obtains $d_{11} = \left(1 - \frac{\lambda_1}{2}\right) \cdot q_1$. In order to become independent of the first generation quantisation step-size q_1, the value of d_{11} can be set to $d_{11} = \left(1 - \frac{\lambda_1}{2}\right) \cdot q_{min}$ where q_{min} is the smallest possible step-size that complies with eq. (2).

Therefore, the pdf is modelled as

$$p(x) = \beta \cdot p_L(x) = \beta \cdot \frac{\alpha}{2} \cdot e^{-\alpha|x|} \qquad ,|x| \geq d_{11} \; . \tag{41}$$

The parameter β in eq. (41) takes into account that the integral $\int p(x)dx$ is smaller than one when calculated only over the range $|x| \geq d_{11}$. The value of β can be used to match the condition

$$\int_{|x| \geq d_{11}} p(x)dx = 1 - \int_{|x| \leq d_{11}} p(x)dx = 1 - \sum_{l < l_0} \frac{n_l}{N} = 1 - f_0 . \qquad (42)$$

The right hand side of (42) is determined by the measured relative frequencies $\frac{n_l}{N}$ of absolute values $|x| = l$ that are commonly mapped onto $y_1 = 0$ by the first generation quantiser. The latter is indicated in eq. (42) by the index l_0 which depends on the decision level d_{11}. Thus, instead of calculating the sum

$$f_0 = \sum_{l < l_0} \frac{n_l}{N}$$

that appears in the right hand side of (42) one can also measure the relative frequency f_0 of the event $y_1 = 0$.

It is important to notice that β is not needed to apply the parametric model of eq. (41) to both the mse and the map cost function. During the calculation of the local centroids with eqs. (22) and (23), the value of β is cancelled out as it appears in the top and in the bottom line in eq. (23). Cancellation also occurs for the map cost function during the calculation of the a-posteriori probabilities of eqs. (36) and (37). Hence, only the value of α is needed in the transcoder. As a consequence, eq. (41) can provide a better fit for the curves shown in Fig. 11 than eq. (40), resulting in a more accurate estimate of the parameter α with no increase of side information.

The probabilities P_{1l} can be evaluated for the parametric description of the decision levels of eq. (5) by inserting the parametric model of eq. (41) in eq. (22). After a straightforward calculation that also takes into account condition (42) one obtains

$$P_{1l} = \frac{1}{2} \cdot (1 - f_0) \cdot (1 - z) \cdot \frac{z^{(|l| - 1)}}{1 - z^L} \qquad ,1 \leq |l| \leq L \qquad ,z = e^{-(\alpha \cdot q_1)} . \qquad (43)$$

We recall that eq. (43) specifies the probability that one observes a first generation coefficient $y_1 = r_{1l} = l \cdot q_1$. It is interesting to notice that the first generation quantiser parameter λ_1 is not needed in eq. (43) but only the step-size q_1. Due to the symmetry assumption of the parametric model (41) one gets $P_{1l} = P_{1(-l)}$ as a outcome of (43). For $l = 0$ the probability is set to $P_{10} = f_0$. The largest index in (43), i.e. L, can be set to the largest possible value $L = 1024$

for AC-coefficients or it may be set more accurately to the largest value that is actually encountered in the transcoder. While eq. (43) forms the basis to estimate the parameter α from the first generation coefficients y_1, we need a corresponding equation to estimate α from the original coefficients x. As already mentioned, the original dct-coefficients result as a special case for the step-size $q_1 = 1$, i.e. $x = y_1 = r_{1l} = l \cdot 1$. When combined with condition (42), eq.(43) has to be modified as follows,

$$P_{1l} = \frac{1}{2} \cdot (1 - f_0) \cdot (1 - z) \cdot \frac{z^{(|l| - l_0)}}{1 - z^{(L - l_0 + 1)}} \qquad ,1 \le l_0 \le |l| \le L \qquad ,z = e^{-\alpha}. \tag{44}$$

As in the previous case, the probabilities for indices in the range $|l| < l_0$ that are not defined by eq. (44) can be set to the value of the corresponding relative frequencies of the original dct-coefficients x.

We first concentrate on estimating the parameter α according with eq. (43). Ideally, the model probabilities P_{1l} of eq. (43) should coincide with the relative frequencies f_l of the first generation coefficients $y_1 = r_{1l} = l \cdot q_1$ that can be measured in the transcoder. However, there is only one free parameter in the model, i.e. α. Therefore, $P_{1l} = f_l$ is in general not achievable for each index l. A coding argument can be used as a guideline for adjusting α. If the parametric model of (43) were applied to encode the first generation coefficients y_1, then the minimum average codeword length cwl would be

$$cwl = \sum_l -f_l \cdot \log P_{1l} \ge \sum_l -f_l \cdot \log f_l = ent, \tag{45}$$

cwl is given in the unit 'bit per coefficient' if the logarithm is taken relative to the base of 2 in eq. (45). The right hand side of (45) specifies the (first order) source entropy ent which is the lower bound for cwl that can only be reached if $P_{1l} = f_l$ holds, throughout. The goal is now to adjust the parameter α of the probabilities P_{1l} such that the resulting cwl is minimised and is as close as possible to the source entropy ent. It is further interesting to notice that cwl can also be written as

$$cwl = \sum_l -f_l \cdot \log P_{1l} = -\frac{1}{N} \cdot \log P(y_{11}, y_{12}, \ldots, y_{1N}), \tag{46}$$

where $P(y_{11}, y_{12}, \ldots, y_{1N})$ specifies the joint probability that results from the parametric model (43) for all first generation coefficients y_1 arranged in some scan order $y_{11}, y_{12}, \ldots, y_{1N}$. As a consequence of (46), the minimisation of cwl by adjusting α coincides with the maximisation of $P(y_{11}, y_{12}, \ldots, y_{1N})$ for the observed coefficients $y_{11}, y_{12}, \ldots, y_{1N}$. Thus we see that the parameter α is determined by a maximum likelihood (ML) estimation.

The ML-estimate for α can be calculated with $\frac{\partial}{\partial \alpha} cwl = 0$. In order to simplify this calculation, eq. (46) is evaluated by inserting the parametric model for P_{1l} that results when the largest index L in eq. (43) tends to infinity. This is justified by the fact that in practice L is very large, e.g. $L = 1024$. For $L \to \infty$, after differentiating and re-ordering one gets the equation

$$\frac{\partial}{\partial \alpha} cwl = q_1 \cdot \left[\dot{l} - \frac{(1-f_0)}{1-z} \right] = 0 \ . \tag{47}$$

In eq. (47) the value of α is indirectly given by $z = e^{-(\alpha \cdot q_1)}$, see also eq. (43), and $\dot{l}$ specifies the measured average value of the absolute first generation indices $|l|$, i.e.

$$\dot{l} = \sum_l |l| \cdot f_l . \tag{48}$$

From eq. (47) one obtains the ML-estimate for α,

$$z = e^{-(\alpha \cdot q_1)} = 1 - \frac{1-f_0}{\dot{l}} . \tag{49}$$

Note that only the mean value $\dot{l}$ of eq. (48) and the relative frequency f_0 of the event $y_1 = 0$ have to be measured from the first generation coefficients to determine the ML-estimate of (49).

One obtains a corresponding result if the parameter α is not estimated from the first generation coefficients y_1 but from the original dct-coefficients x. In this case eq. (44) instead of eq. (43) is used to derive the ML-estimate

$$z = e^{-\alpha} = \frac{\dot{l} - (1-f_0) \cdot l_0}{\dot{l} - (1-f_0) \cdot (l_0 - 1)} , \tag{50}$$

where the mean value $\dot{l}$ is given by

$$\dot{l} = \sum_{l \geq l_0} l \cdot \frac{n_l}{N} . \tag{51}$$

As introduced earlier the ratios $\frac{n_l}{N}$ in eq. (51) specify the relative frequencies of the abolute values $|x| = l$, see also Fig. 11. The ML estimation of eqs. (50) and (51) coincides with the ML estimation of eqs. (48) and (49) in the special case of $q_1 = l_0 = 1$. In general, one obtains different estimation values for the parameter α. As more information is provided by the orig-

inal dct-coefficients, the ML-estimation with (50) and (51) is more accurate, however, the parameter α has then to be signalled as additional side information. In this case it may be more convenient to signal $z = e^{-\alpha}$ rather than α because the z-values have a normalised amplitude range, i.e. $0 \le z \le 1$, so that each z-value can be rounded to a fractional binary number of e.g. 8 bit length. In contrast to (50) and (51), the ML-estimation with (48) and (49) requires no additional side information that has to be sent to the transcoder.

As an example, the parameter α has been estimated for each curve of Fig. 11 from the original dct-coefficients x by applying eqs. (50) and (51). In order to determine the value of l_0 the decision level d_{11} has been set according to

$$d_{11} = \left(1 - \frac{\lambda_1}{2}\right) \cdot q_1 = \left(1 - \frac{1}{2}\right) \cdot \frac{w_1}{16} \cdot qscale_{min} = \frac{w_1}{16}, \tag{52}$$

resulting in

$$l_0 = \lceil d_{11} \rceil = \left\lceil \frac{w_1}{16} \right\rceil, \tag{53}$$

where the function $\lceil a \rceil$ rounds the given argument a up to the nearest integer. The resulting ML-estimate of $z = e^{-\alpha}$ can be used for all $qscale_1$-values that are larger or equal to $qscale_{min} = 2$. According to eq. (53) the value of l_0 depends only on the visual weight w_1 that changes with the horizontal and vertical frequency index of the AC-coefficients. The weighting matrix of MPEG-2 test model TM5 has been used. Additionally, the estimated z-values have been rounded to fractional binary numbers of 8 bit length. Figs. 12a - 12c show the resulting model probabilities of eq. (44) in comparison to the measured histograms of Fig. 11. Note that in Figs. 12a-12c the model probabilities for the absolute amplitude levels are shown, i.e. $P_{1|l|} = P_{1l} + P_{1(-l)}$ for $|l| \ge 1$.

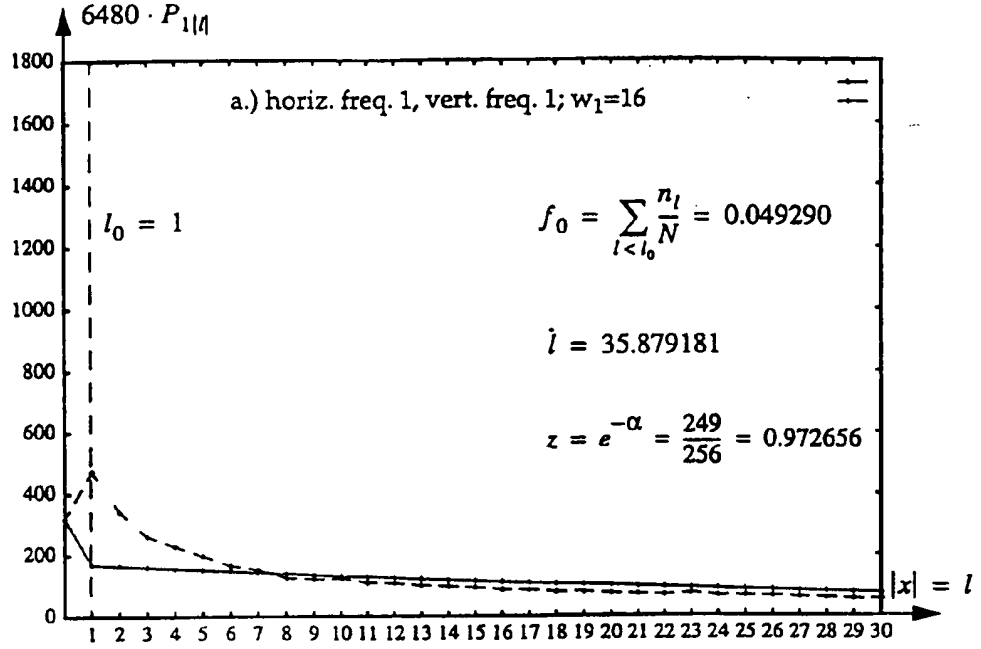


Fig. 12 a: Model probabilities acc. to eq. (44) (solid line) and histogram (dashed line) for absolute amplitude levels of original dct-coefficients, test signal 'mobile'

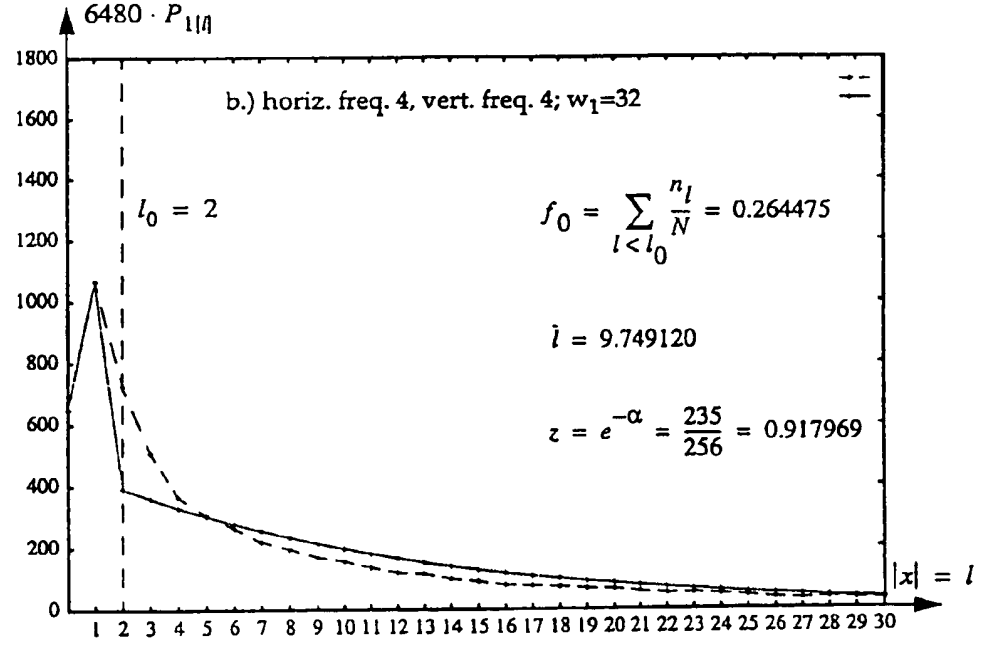


Fig. 12 b: Model probabilities acc. to eq. (44) (solid line) and histogram (dashed line) for absolute amplitude levels of original dct-coefficients, test signal 'mobile'

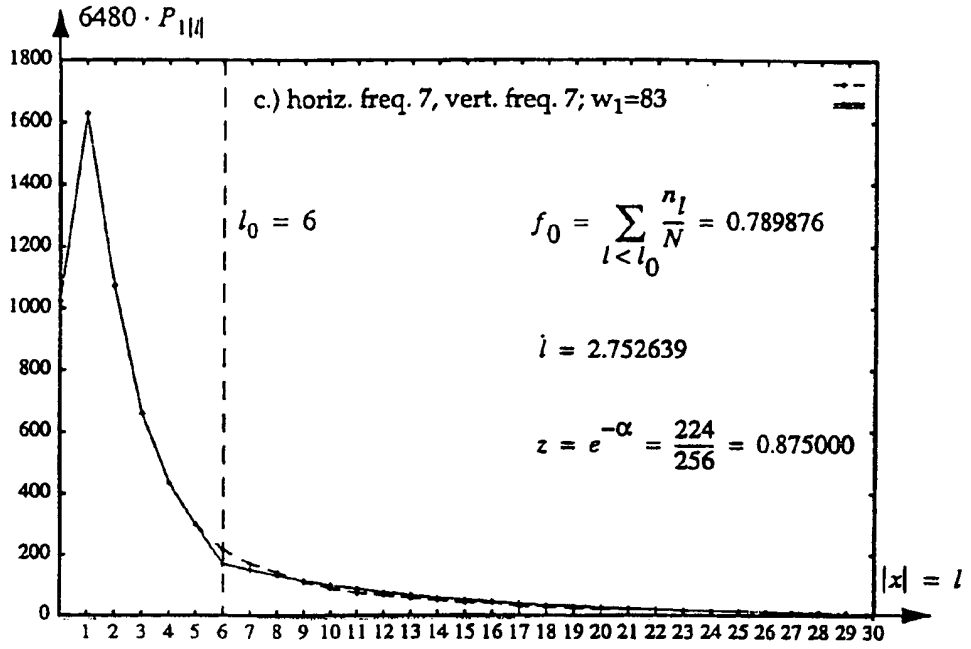


Fig. 12 c: Model probabilities acc. to eq. (44) (solid line) and histogram (dashed line) for absolute amplitude levels of original dct-coefficients, test signal 'mobile'

6. Evaluation of the mse and the map cost function for the parametric model

In this section both the mse and the map cost function are evaluated for the parametric model of the pdf $p(x)$ of eq. (41). The model parameter α can be estimated as described in the previous section. In order to ease the comparison between the mse and the map cost function, the resulting decision rules of eq. (34) and (38) are considered, respectively.

Firstly, the mse decision rule of eq. (34) is evaluated. The pdf $p(x)$ determines the parameter μ_{1l} that is related to the local centroid c_{1l} according to eq. (30). The local centroid c_{1l} can be calculated for the decision levels of eqs. (25) and (26) by inserting the parametric model of eq. (41) in eqs. (22) and (23). Without loss of generality a positive index $l \geq 1$ can be considered, the result can be mirrored for the corresponding negative index $-l$. The calculus then yields

$$c_{1l} = d_{1l} + \frac{1}{\alpha} \cdot \frac{1 - (1 + \alpha \cdot q_1) \cdot e^{-(\alpha \cdot q_1)}}{1 - e^{-(\alpha \cdot q_1)}}. \qquad (54)$$

From eq. (30) one gets $c_{1l} = d_{1l} + \frac{\mu_{1l}}{2} \cdot q_1$ and by comparison with eq. (54)

$$\frac{\mu_{1l}}{2} = \frac{1}{\alpha \cdot q_1} \cdot \frac{1 - (1 + \alpha \cdot q_1) \cdot e^{-(\alpha \cdot q_1)}}{1 - e^{-(\alpha \cdot q_1)}}. \tag{55}$$

It follows from eq. (55) that the local centroid parameter μ_{1l} does not depend on the index l that is selected in the transcoder by the actual first generation coefficient $y_1 = r_{1l} = l \cdot q_1$, rather μ_{1l} depends only on the estimated parameter α and the step-size q_1 that is decoded from the first generation bit stream. Hence, the Laplacian-like pdf defined by the parametric model of eq. (41) is another case that results in a constant parameter μ_{1l} apart from the special cases already discussed in the last paragraph of Section 4.1.1. Consequently, the parametric model of eq. (41) also reduces the complexity for implementing the mse decision rule in the transcoder.

Secondly, the parametric model is evaluated for the map decision rule of eq. (38). After the computation of the a-posteriori probabilities of eqs. (36) and (37) for the decision levels of eqs. (25), (26) and the model pdf of eq. (41), the map decision rule of eq. (38) can be rewritten in a similar form to the mse decision rule of eq. (34),

$$y_2 = Q_2(y_1 = r_{1l}) = \begin{cases} r_{2(L-1)} & \text{if } \frac{d_{L,ref} - d_{1l}}{q_1} > \frac{v_1}{2} \\ r_{2L} & \text{if } \frac{d_{L,ref} - d_{1l}}{q_1} < \frac{v_1}{2} \end{cases}. \tag{56}$$

The map-threshold in eq. (56) is given by

$$\frac{v_1}{2} = \frac{1}{\alpha \cdot q_1} \cdot \ln\left(\frac{2}{1 + e^{-(\alpha \cdot q_1)}}\right), \tag{57}$$

where $\ln(a)$ returns the natural logarithm, i.e. logarithm relative to the base e, of the argument a.

Thus we see that the map decision rule can be implemented in essentially the same way as the mse decision rule for the parametric model of eq. (41). The main differences are that the mse-threshold is defined by eq. (55) but the map-threshold by eq. (57), and that the decision levels $d_{L,\,ref}$ are fixed in the mse case but can vary depending on the reference quantiser $Q_{2,\,ref}$ in the map case as discussed earlier in Section 4.2.

The same estimated parameter α can be applied to both the mse and the map cost functions. Therefore, the transcoder has the option to switch locally e.g. on a 8x8 block basis, between the mse and the map cost function with no further increase of additional side information.

7. Experimental results

In order to verify the theoretical results derived in the previous sections, the transcoding set up of Figs. 3 and 4 has been simulated for ten consecutive frames of the CCIR-601 [CCIR-601] formatted test signal 'mobile'. As the MPEG-2 test model TM5 [TM5-93] is a widely acknowledged reference for a MPEG-2 standalone encoder, the first generation quantiser Q_1 has been fixed to the TM5 quantiser characteristic throughout in the experiments.

The transcoder's quantiser characteristic Q_2 has been set to TM5 in the first experiment. As an example, the corresponding qscale value has been fixed to $qscale_2 = 32$, which corresponds approximately to the adjustment in I-frames that is used for a 3 Mbit/s simulation including P- and B-frames. According to the *linear qscale table* of MPEG-2, the qscale value of the TM5-quantiser used in the first generation encoder has been varied in the range $qscale_1 = 2, 4, 6, 8, \ldots, 32$. Thus, transcoding is simulated for different first generation bit rates and a fixed target bit rate for the second generation. The resulting PSNR values for the image signal s_2 reconstructed from the second generation bit stream b_2 are shown in Fig. 13a as a function of $qscale_1$. For reference, the solid line shows the resulting PSNR value of a standalone TM5-encoder that by-passes the first generation and directly encodes the original signal with a qscale value of 32. The Peak-Signal-to-Noise-Ratio is related to the mean-squared-error between the original and the second generation signal as $PSNR = 10 \cdot \log_{10}[255^2/mse]$. In addition, Fig 13b conveys the bit amount needed to encode the second generation AC-coefficients with the MPEG-2 *intra vlc table*.

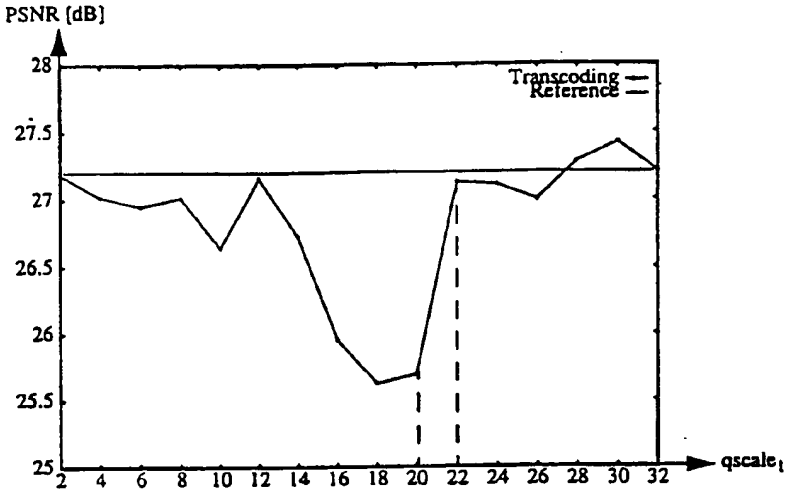


Fig. 13a: PSNR of 2nd generation signal as a function of $qscale_1$ for fixed $qscale_2 = 32$, TM5, test signal 'mobile'

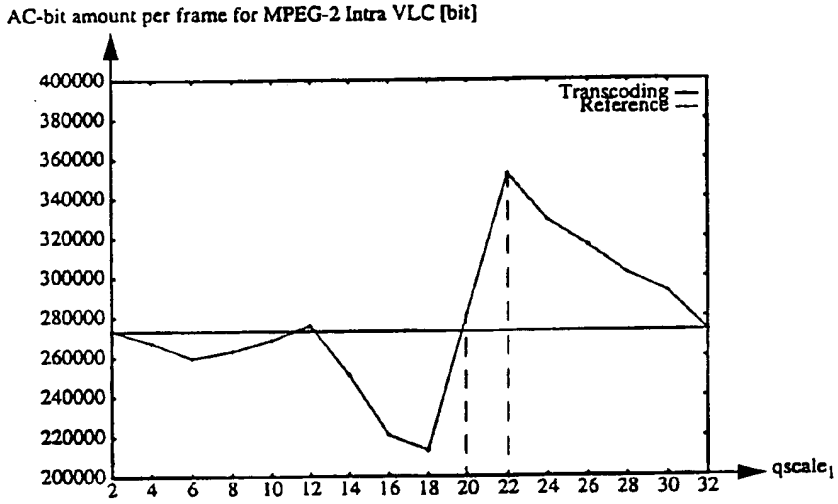


Fig. 13b: AC-bit amount of 2nd generation signal as a function of $qscale_1$ for fixed $qscale_2 = 32$, TM5, test signal 'mobile'

The PSNR values drop dramatically by more than 1 dB for medium values of $qscale_1$ in the range of 16-20 when compared to the reference. At the same time, the resulting bit amount changes considerably. In this case the resulting ratio $qscale_2/qscale_1 = q_2/q_1$ is around 2, which is unfavourable for transcoding as explained during the discussion of eq. (17) in Section 3. As an example, for $qscale_1 = 20$, the PSNR value is about 1.5 dB below the reference line while

the resulting bit amount is approximately (within 3%) the same for the reference and the transcoded signal, see Figs. 13a, 13b. When $qscale_1$ is increased to 22, the PSNR value recovers and is only 0.1 dB below the reference line, however, now the bit amount dramatically exceeds the reference line by about 29%. These considerable changes between consecutive $qscale_1$ values would have the most unpleasant effect on any rate control scheme that is used in the transcoder, thus resulting in a poor picture quality for the second generation. The PSNR value recovers when the $qscale_1$ value approaches $qscale_2$, but the resulting bit amount is rather high. For $qscale_1 = qscale_2 = 32$, transparency is achieved, i.e. the first and the second generation bit stream are identical. The TM5-quantiser provides the best results for small values of $qscale_1$ in the range 2-8. Clearly, the first generation dct-coefficients contain less quantisation noise when $qscale_1$ is small, and as a consequence, the difference between the second generation and the reference signal also becomes small. Similar results have been obtained for other test signals [OW-BBC-1].

In addition to the TM5 quantiser characteristic, the transcoder has the option to apply the mse or the map cost function. The decision to select either of them may be based on the rate-distortion performance. While the resulting PSNR/mse-values can be estimated and compared in the transcoder by using the parametric model of eq. (41) for the pdf $p(x)$, it would be rather complex if the resulting bit amount had to be computed for each cost function by forming the MPEG-2 compatible two-dimensional *(runlength,amplitude)*-events of each 8x8 block and by looking up each codeword length from the *intra vlc table*. It is less complex for the transcoder to calculate the first order source entropy, as in eq. (45), because only a histogram count for the second generation coefficients is then needed. Fig. 14a shows the bit amount for the MPEG-2 *intra vlc* of Fig. 13b scaled down for 8x8 blocks in comparison to the source entropy summed up for all 63 AC-coefficients. It can be seen that the two curves shown in Fig. 14a have essentially the same shape. This is confirmed if one calculates the ratio between them for each $qscale_1$ value, see Fig 14b. The ratio is almost constant, i.e. 0.84. Thus, a comparison of the mse and the map cost function and the TM5 performance in terms of bit rate can also be carried out on the basis of the source entropy rather than by applying the MPEG-2 *intra vlc*. Another interesting result of Fig. 14b is that up to approx. 16% of the bit rate could be saved by using the AC-entropy source model instead of the MPEG two-dimensional *(runlength,amplitude)* model. However, this result cannot be exploited for MPEG-2 compatible bit streams.

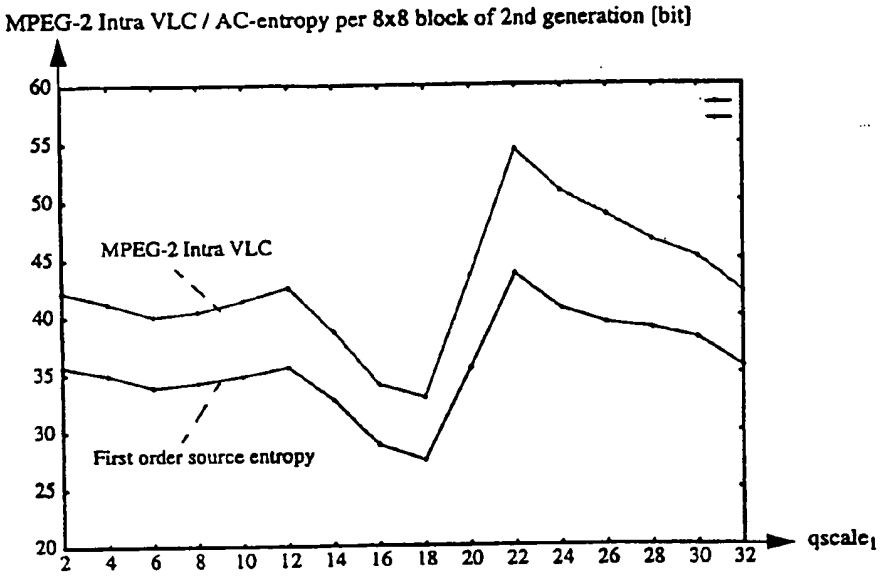


Fig. 14 a: Bit amount of 2nd generation AC coefficients: MPEG-2 Intra VLC vs. first order source entropy, test signal 'mobile'

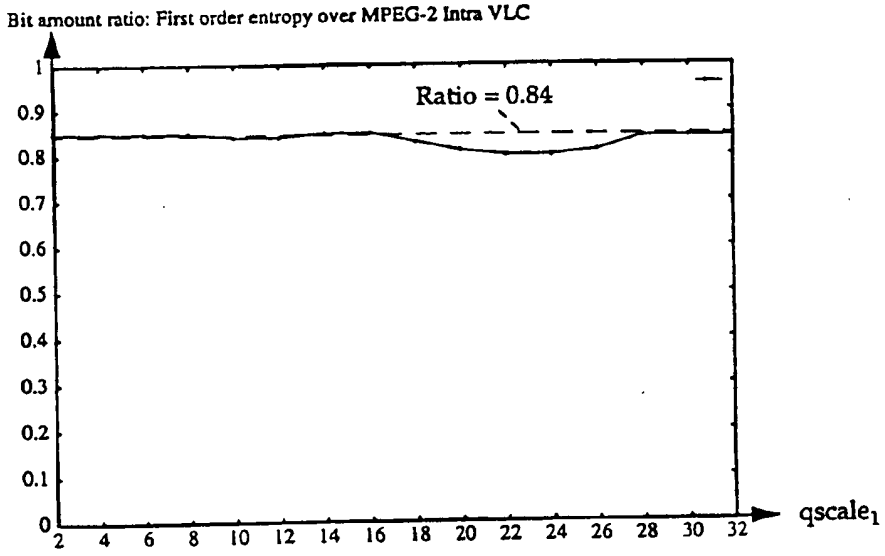


Fig. 14 b: Bit amount ratio: First order source entropy over MPEG-2 Intra VLC (solid line), ratio =0.84 (dashed line), test signal 'mobile'

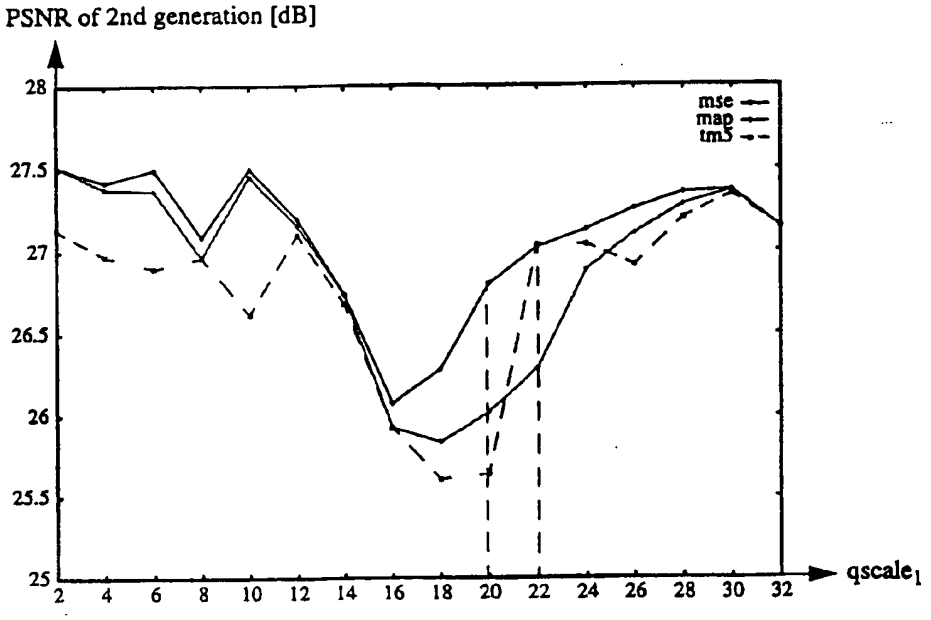


Fig. 15a: PSNR of 2nd generation signal as a function of $qscale_1$ for fixed $qscale_2 = 32$; mse(upper line), map/$\lambda_{2,ref} = 0.90$(middle line), TM5(dashed line), test signal 'mobile'

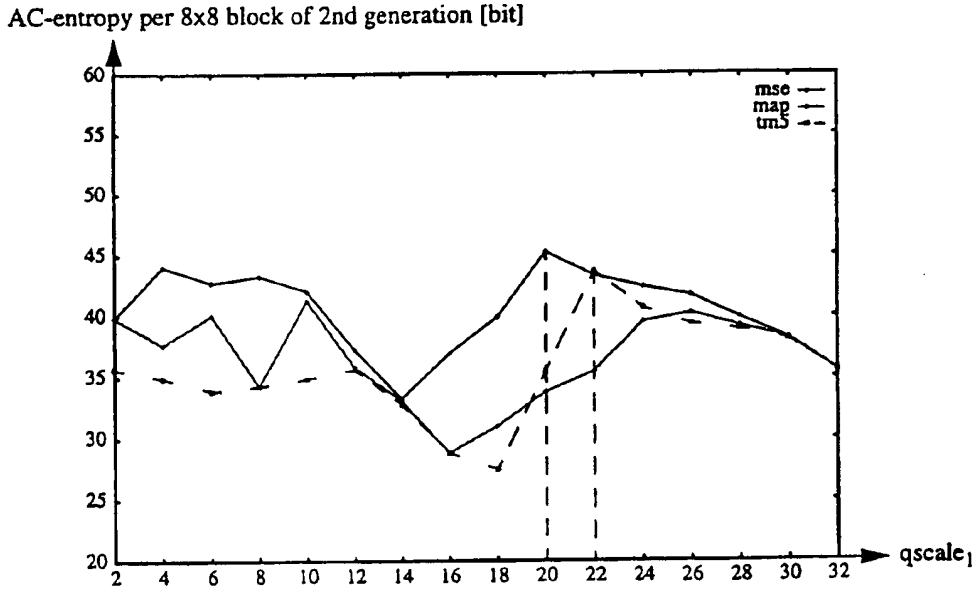


Fig. 15b: AC-entropy of 2nd generation signal as a function of $qscale_1$ for fixed $qscale_2 = 32$; mse(upper line), map/$\lambda_{2,ref} = 0.90$(middle line), TM5(dashed line), test signal 'mobile'

The mse and the map cost function have been simulated in the next experiment. The reference quantiser $Q_{2,\,ref}$ of the map cost function complies to eqs. (3) and (4), as an example the decison level parameter has been set to $\lambda_{2,\,ref} = 0.90$. The pdf parameter α has been estimated for each AC-frequency index as described in the example given in Section 5, see Figs. 12a-12c, with respect to the original dct-coefficients, resulting in $63 \times 8 = 504$ bit additional side information. We recall that these α-values do not depend on the actual $qscale_1$ value as they are estimated according to eqs. (52) and (53). Thus, the additional side information of 504 bit is the same for all first generation bit streams. A more accurate estimation of α is possible if the actual $qscale_1$ value is inserted in eq. (52) instead of $qscale_{min} = 2$; however, the additional side information would then depend on the first generation encoding process.

A comparison of the cases where either the TM5 quantiser, the mse or the map cost function is used in the transcoder is shown in Figs. 15a and 15b. The mse cost function achieves the largest PSNR values, see Fig. 15a. For example in the case of $qscale_1 = 20$, the PSNR value can be increased by approx. 1.1 dB from approx. 25.6 dB for the TM5-quantiser to approx. 26.7 dB for the mse cost function. However, the mse cost function comes at a price as the entropy of the AC-coefficients is significantly increased, see Fig. 15b. The mse cost function may therefore not be applicable throughout. However, the mse cost function can be used locally to transcode blocks with critical image content. As a further option, the mse cost function can only be applied to AC-coefficients with low frequency index because the human visual system is in general more sensitive to quantisation noise that is added to low frequencies.

The map cost function is more suitable in a rate-distortion sense compared with the mse cost function. For the critical case of $qscale_1 = 20$, the AC-entropy is approximately 4.7% smaller and at the same time the PSNR value is approximately 0.4 dB larger in comparison to TM5. Figs. 15a and 15b show just one example for the map cost function. As the map cost function is governed by the parameter $\lambda_{2,\,ref}$, a set of rate-distortion characteristics can be generated by varying $\lambda_{2,\,ref}$, thus allowing a smooth transition between the TM5-quantiser performance and the mse cost function in terms of PSNR and corresponding bit rate. The results for different values of $\lambda_{2,\,ref}$ in Figs. 16a and 16b show a monotonic behaviour: when $\lambda_{2,\,ref}$ is increased, both the PSNR value and the AC-entropy are increased, too.

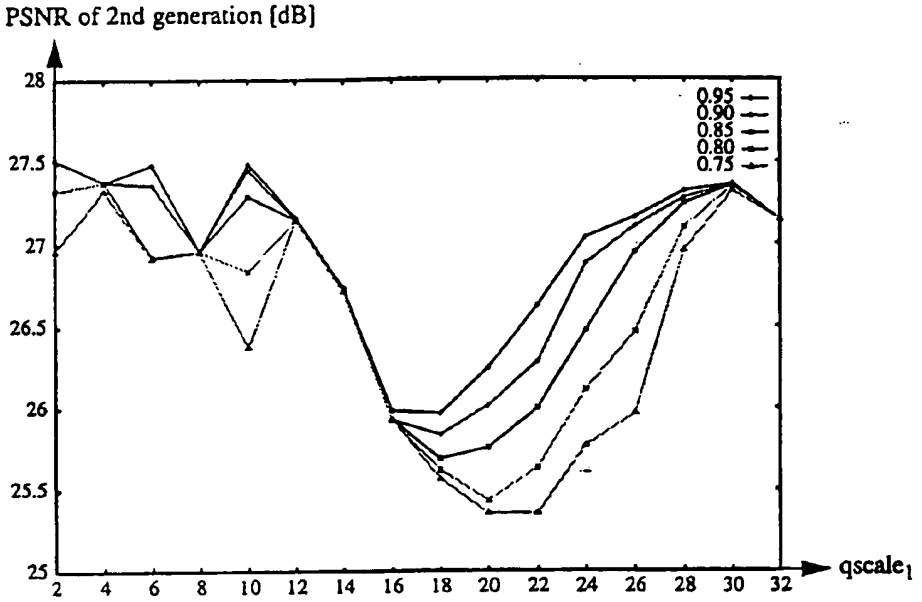


Fig. 16a: PSNR of 2nd generation signal as a function of $qscale_1$ for fixed $qscale_2 = 32$; map cost function for different values of $\lambda_{2,ref}$, test signal 'mobile'
$\lambda_{2,ref} = 0.95$ (upper line), 0.90, 0.85, 0.80, 0.75 (lower line)

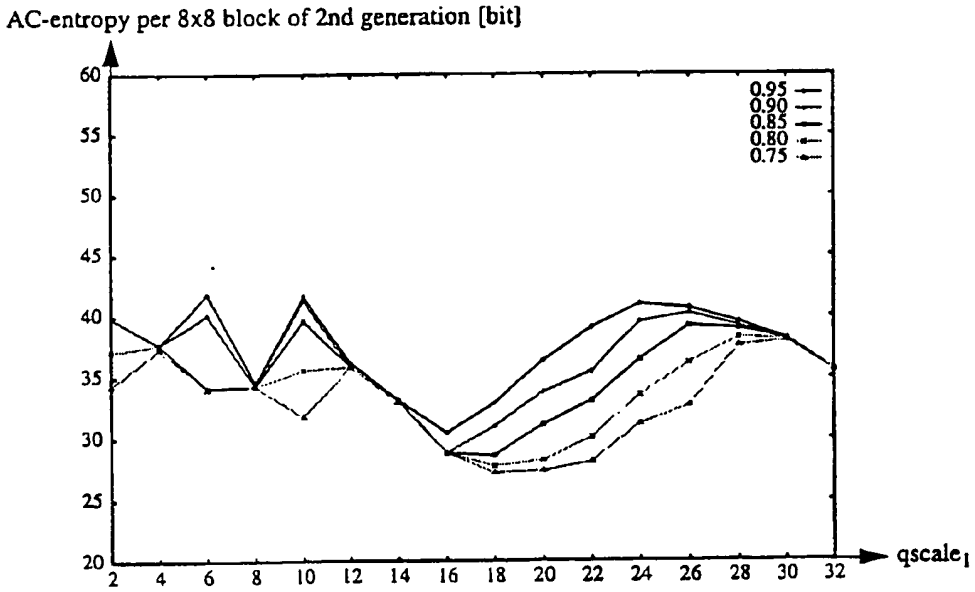


Fig. 16b: AC-entropy of 2nd generation signal as a function of $qscale_1$ for fixed $qscale_2 = 32$; map cost function for different values of $\lambda_{2,ref}$, test signal 'mobile'
$\lambda_{2,ref} = 0.95$ (upper line), 0.90, 0.85, 0.80, 0.75 (lower line)

In the last experiment the impact of estimating the pdf parameter α not from the original dct-coefficients but from the first generation coefficients is investigated. No additional side information is necessary if α is estimated from the first generation coeffcients. A comparison between the two methods is shown in Figs. 17a and 17b for the mse cost function. While the PSNR/mse performance in Fig. 17a is almost identical with and without additional side information, one can see from Fig. 17b that the resulting bit amount is significantly higher in some cases without side information. Thus, not very surprisingly, the results are in favour of sending additional side information. This conclusion is not necessarily true for the map cost function as shown for the example of $\lambda_{2,\,ref} = 0.90$ in Figs. 18a and 18b. Similar to the mse cost function, the AC-entropy is significantly higher for some $qscale_1$ values when no side information is sent (Fig. 18b), however, the resulting PSNR values also exceed the corresponding PSNR values for the case of additional side information. Thus, a good performance can be achieved even when no additional side information is sent to the transcoder.

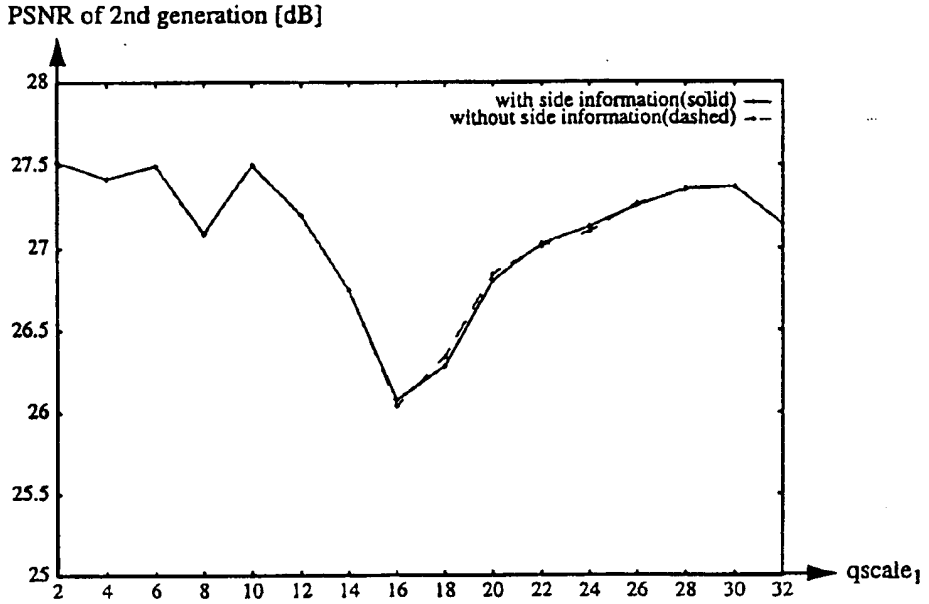


Fig. 17a: PSNR of 2nd generation signal as a function of $qscale_1$ for fixed $qscale_2 = 32$; mse cost function with and without additional side information, test signal 'mobile'

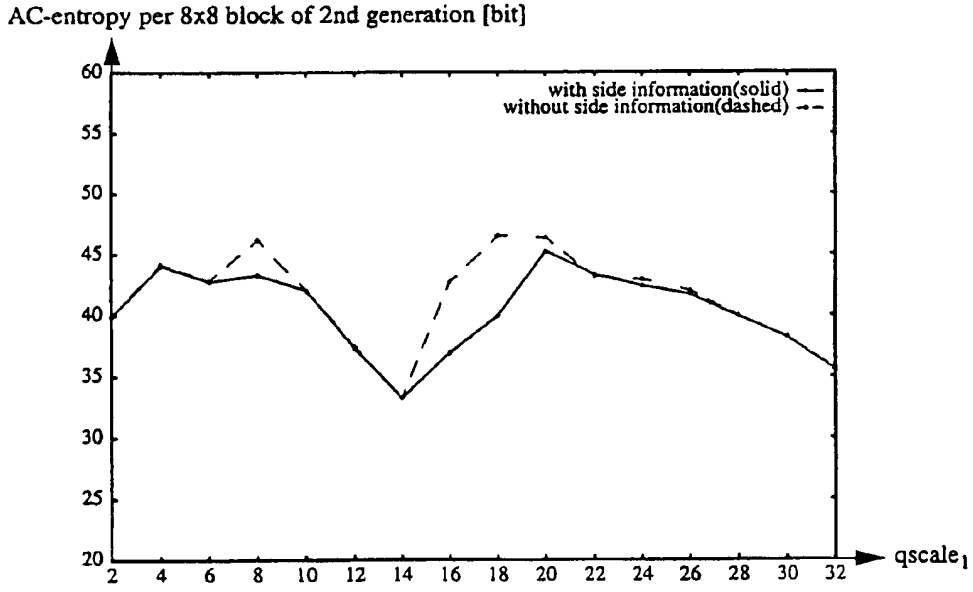


Fig. 17b: AC-entropy of 2nd generation signal as a function of $qscale_1$ for fixed $qscale_2 = 32$; mse cost function with and without additional side information, test signal 'mobile'

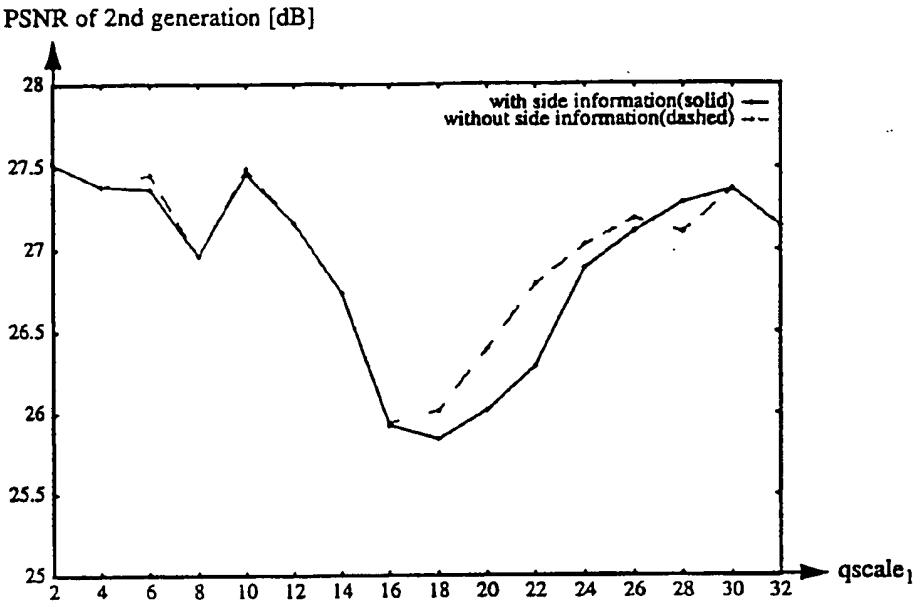


Fig. 18a: PSNR of 2nd generation signal as a function of $qscale_1$ for fixed $qscale_2 = 32$; **map** cost function with and without additional side information, $\lambda_{2,ref} = 0.90$, test signal 'mobile'

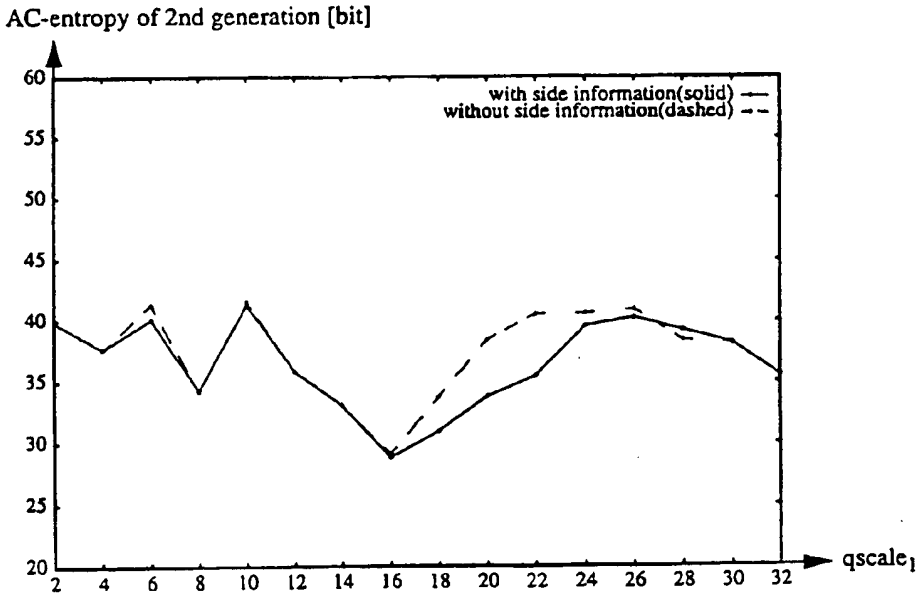


Fig. 18b: AC-entropy of 2nd generation signal as a function of $qscale_1$ for fixed $qscale_2 = 32$; **map** cost function with and without additional side information, $\lambda_{2,ref} = 0.90$, test signal 'mobile'

8. Conclusions and future work

This paper discusses transcoding of MPEG-2 intra frames. A theoretical analysis of the transcoding problem is carried out with emphasis on designing the quantiser characteristic of the transcoder. The comparison with a reference quantiser of a standalone encoder shows that transcoding only results in an equivalent overall quantiser characteristic under special conditions. This is indicated by the ratio of the first and the second generation quantisation step-sizes. Only for certain ratios is the mapping of the original dct-coefficients onto the second generation coefficients the same for the reference quantiser and the overall quantiser characteristic that results from transcoding. In general for an arbitrary ratio, the standalone reference quantiser characteristic cannot be achieved during transcoding. As a consequence, there is a loss due to transcoding when the mean-squared-error (mse) or the rate-distortion performance is considered. In order to minimise this loss, the degree of freedom that lies in selecting the decision levels for an MPEG-2 compatible quantiser can be exploited.

Two approaches for the adjustment of the decision levels in the transcoder are proposed, explained and compared. Firstly, the minimisation of the mse cost function is taken as a guideline for adjusting the decision levels. The objective of the mse cost function is to give the smallest mse values between the original and the second generation coefficients. The resulting mse decision rule can be implemented in the transcoder in essentially the same way as the first generation quantiser characteristic. However, the mse cost function comes at a price as no attention is paid to the bit rate needed to encode the second generation coefficients. Therefore, the maximum a-posteriori (map) cost function is additionally introduced. The map cost function is more suitable in a rate-distortion sense than the mse cost function. While the standalone reference quantiser is fixed for the mse cost function, the map cost function has the additional freedom to choose the standalone reference quantiser. Thus, by changing the reference quantiser, a set of rate-distortion characteristics can be generated with the map cost function. Interestingly, the map cost function is again given by a mean-squared-error (mse); however in contrast to the mse cost function, the mse between the output of the standalone reference quantiser and the second generation coefficients is minimised.

The statistical distribution of the original dct-coefficients is needed to apply both the mse and the map cost function in the transcoder. A parametric model based on a Laplacian probability density function (pdf) is proposed. One parameter for each frequency component allows adaptation to the actual signal statistic. This parameter can be estimated either in the transcoder from the first generation coefficients or from the original dct-coefficients. In general, the latter results in a more accurate estimate, however, the parameter has then to be transmitted as ad-

ditional side information along with the first generation bit stream. For both cases, a maximum likelihood estimation rule is derived. The parametric model is validated with real image data; experimental results confirm that the model pdf is suitable for describing the distribution of the dct-coefficients.

The mse and the map cost function are evaluated for the parametric model. It is shown that the proposed model pdf also simplifies the implementation, and that the resulting mse and map decision rules can then be stated in very similar forms.

Experimental results confirm the effectiveness of the mse and the map cost function. For reference, the quantiser characteristic of the MPEG-2 reference encoder TM5 [TM5-93] has also been used in the transcoder. The results show large changes in terms of PSNR values and bit rate of the second generation coefficients for a ratio around two of the second and first generation quantisation step-sizes. The PSNR value can drop by about 1.5 dB while the bit rate remains constant or, conversely, the PSNR value remains rather constant while the bit rate is increased by almost 30%. This causes problems for a rate controller that is used in the transcoder. The mse cost function achieves the largest PSNR values, resulting in up to 1.1 dB gain compared to TM5. However, the mse cost function also leads to the largest bit rates and may therefore only be applied locally to blocks with critical image content. As a further option, the mse cost function can only be applied to AC-coefficients with low frequency indices because the human visual system is in general more sensitive to quantisation noise that is added to low frequencies.

Experimental results show that in comparison to TM5, the map cost function can lead to a smaller bit rate (4.7 %) and at the same time to a larger PSNR value (0.4 dB) in critical cases, thus resulting in a better rate-distortion performance. By changing the reference quantiser of the map cost function, a set of rate-distortion characteristics can be generated allowing a smooth transition between the rate-distortion performance of the mse cost function and TM5. In the experiments, the class of reference quantisers is spanned by a single parameter; results show a monotonic behaviour in that an increase of the reference quantiser parameter leads to a larger PSNR value and to a larger bit rate.

Experimental results confirm that for the mse cost function the estimation of the pdf model parameter from the original dct-coefficients leads to a better rate-distortion performance than estimating the model parameter from the first generation coefficients. This is not necessarily true for the map cost function, as revealed in one example. Thus, a good transcoding performance can also be achieved when no additional side information is transmitted along with the first generation bit stream.

It is further shown that the first order source entropy of the second generation coefficients can be used to derive an estimate of the bit rate that results from the MPEG-2 *intra vlc* codeword table. This would simplify the computation of the bit rate if the transcoder had to decide upon either the TM5, the mse or the map cost function based on the best rate-distortion performance. The resulting PSNR values can be compared in the transcoder on the basis of the Laplacian model pdf. This could be simplified for the map cost function due to the monotonic behaviour of the rate-distortion performance, e.g. after setting a target bit rate on a frame or block basis, the parameter of the reference quantiser can be increased until the first order source entropy exeeds the target bit rate. The investigation of an 'easy-to-implement' algorithm based on the above rate-distortion considerations is a promising goal of future work. Furthermore, the presented results can be adapted for transcoding of MPEG-2 inter-frames, i.e. P- and B-frames, involving motion compensating prediction. However, the problem of drift [OW-94] [OW-96] between the predictors of the encoder and the decoder has then additionally to be taken into account.

CLAIMS

1. A method for compression encoding of a digital signal, including the steps of conducting a transformation process to generate values and quantising the values through partitioning the amplitude range of a value into a set of adjacent intervals, whereby each interval is mapped onto a respective one of a set of representation levels which are to be variable length coded, such that a bound of each interval is controlled by a parameter λ, characterised in that λ is controlled so as to vary dynamically the bound of each interval with respect to the associated representation level.

2. A method according to Claim 1, wherein each value is arithmetically combined with λ.

3. A method according to Claim 1or Claim 2, wherein λ is a function of the quantity represented by the value.

4. A method according to Claim 3, wherein the transformation is a DCT and λ is a function of horizontal and vertical frequency.

5. A method according to any one of the preceding claims, wherein λ is a function of the quantisation step size.

6. A method according to any one of the preceding claims, wherein λ is a function of the value amplitude.

7. A method according to any one of the preceding claims, wherein the digital signal to be encoded has been subjected to previous encoding and decoding processes and λ is controlled as a function of a parameter in said previous encoding and decoding processes.

8. A method according to Claim 7, having quantisation step size $q = q_2$ and a value of $\lambda = \lambda_2$. in which the value to be quantised has previously been quantised using a quantisation step size $q = q_1$ and a value of $\lambda = \lambda_1$, wherein λ is a function of q_1 and λ_1 .

9. A method according to Claim 7 or Claim 8, wherein λ is a function of λ_{ref}, where λ_{ref} is the value of λ that would have been selected in a method according to Claim 1 operating with a quantisation step size $q = q_2$ upon the value prior to quantisation with the quantisation step size $q = q_1$.

10. A method according to any one of the preceding claims, wherein the quantisation step size **q** is independent of the input value, otherwise than for the zero quantisation level.

11. A (q, λ) quantiser operating on a set of transform coefficients x_k representative of respective frequency indices f_k in which λ is dynamically controlled in dependence upon the values of x_k and f_k

12. A quantiser according to Claim 11, wherein the parameters f_k are frequency indices.

13. A quantiser according to Claim 11 or Claim 12, in which λ is dynamically controlled to minimise a cost function $D + \mu H$ where D is a measure of the distortion introduced by the quantisation in the uncompressed domain and H is a measure of compressed bit rate and μ is a constant determined by the bit rate constraint.

14. In a compression transcoder, operating on a compressed signal quantised in first (q_1, λ_1)-type quantiser, a second (q_2, λ_2)-type quantiser in which the second generation λ_2 value is controlled as a function:

$$\lambda_2 = \lambda_2(f_{hor}, f_{ver}, q_1, \lambda_1, q_2, \lambda_{2,ref}, y_1)$$

15. A compression transcoder according to Claim 14 in which the parameter $\lambda_{2,ref}$, represents a notional reference ($q_2\lambda_{2,ref}$) -type quantiser which bypasses the first generation coding and directly maps an original amplitude onto a second generation reference amplitude.

16. A compression transcoder according to Claim 14 in which the parameter $\lambda_{2,ref}$ is selected empirically.

17. A compression transcoder according to Claim 16 in which the parameter $\lambda_{2,ref}$ is fixed for each frequency.

Fig.1.

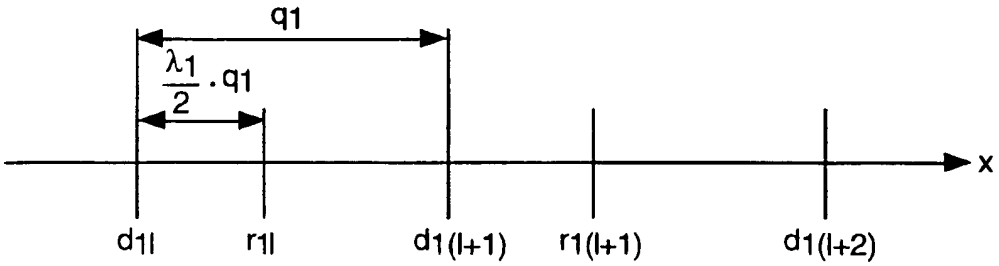

q1
λ1/2 . q1
x
d1l
r1l
d1(l+1)
r1(l+1)
d1(l+2)

Fig.2.

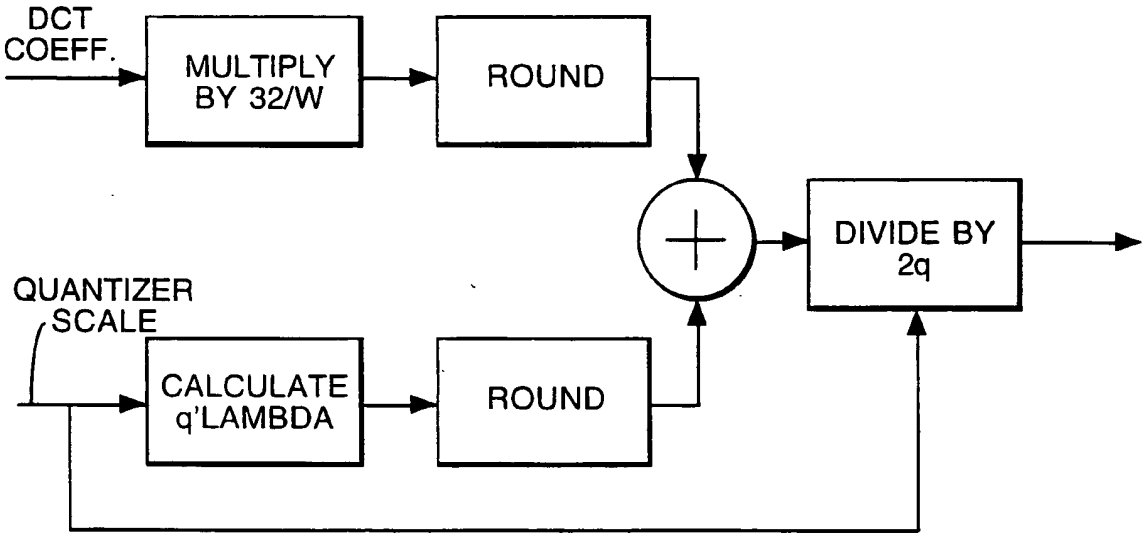

DCT COEFF.
MULTIPLY BY 32/W
ROUND
DIVIDE BY 2q
QUANTIZER SCALE
CALCULATE q'LAMBDA
ROUND

Fig.3.

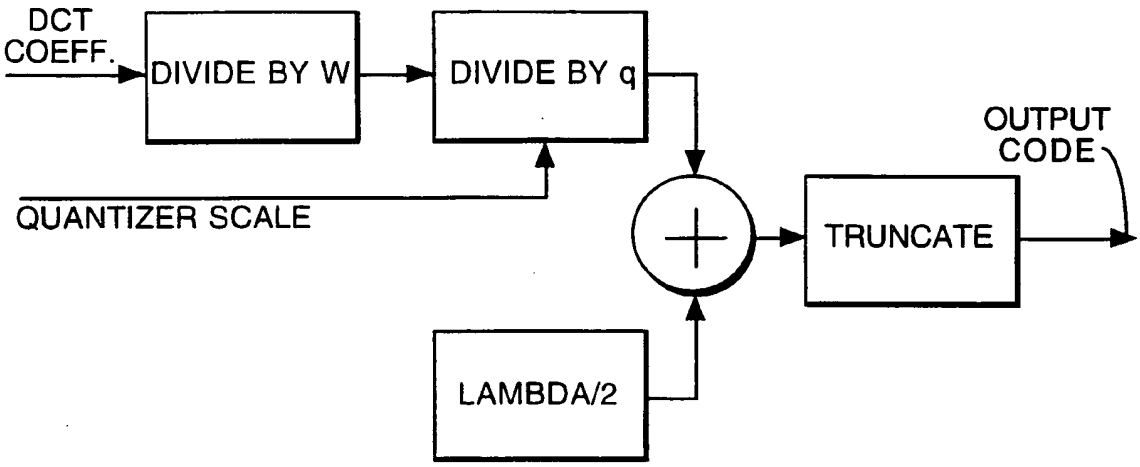

DCT
COEFF.
DIVIDE BY W
DIVIDE BY q
OUTPUT
CODE
QUANTIZER SCALE
TRUNCATE
LAMBDA/2

Fig.4.

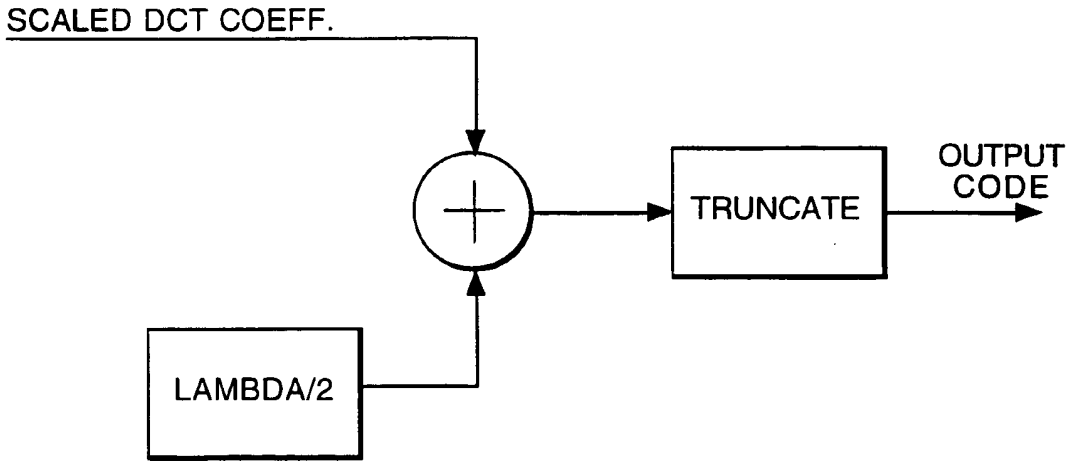

SCALED DCT COEFF.
OUTPUT
CODE
TRUNCATE
LAMBDA/2

Fig.5.

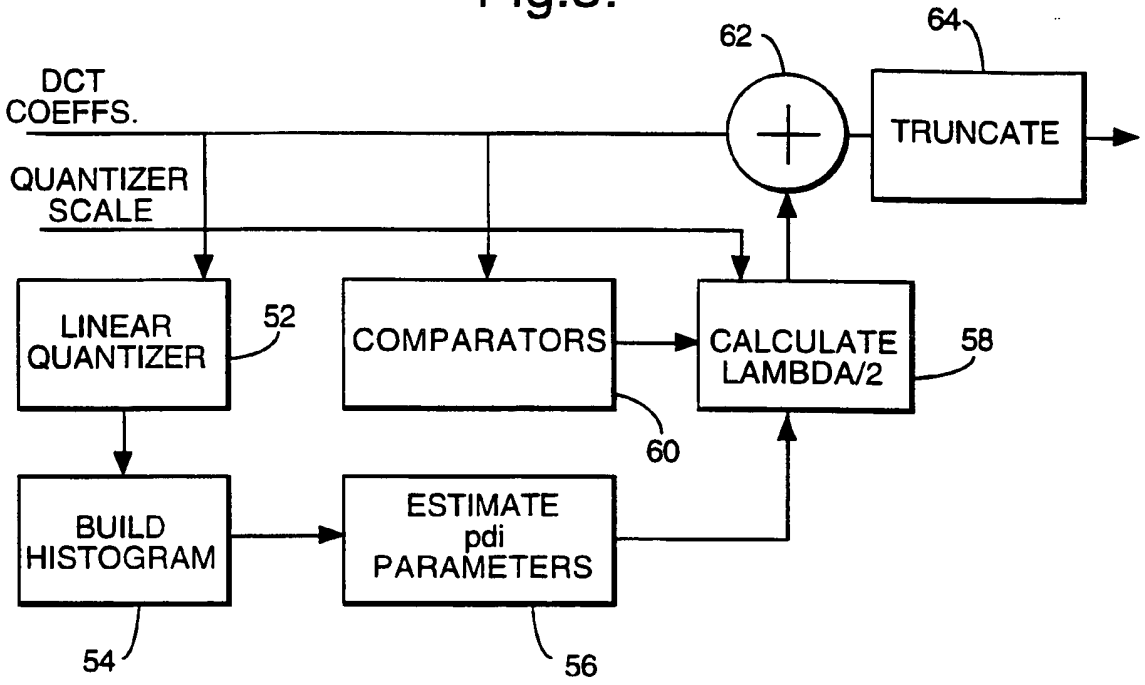
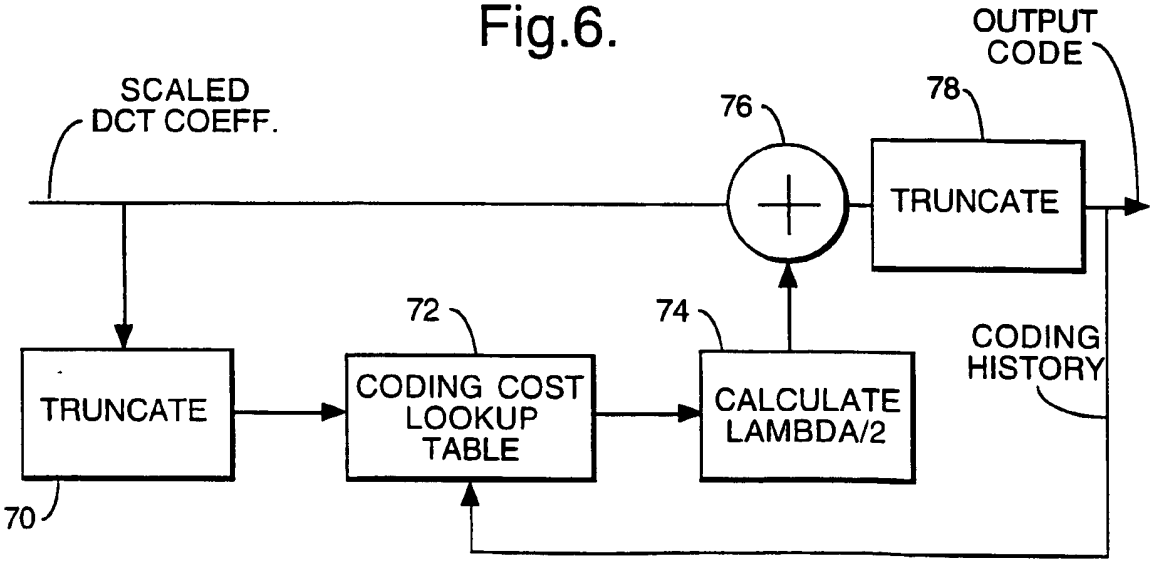

62
64
DCT COEFFS.
TRUNCATE
QUANTIZER SCALE
LINEAR QUANTIZER
52
COMPARATORS
CALCULATE LAMBDA/2
58
60
BUILD HISTOGRAM
ESTIMATE pdi PARAMETERS
54
56

Fig.6.


OUTPUT CODE
SCALED DCT COEFF.
76
78
TRUNCATE
72
74
CODING HISTORY
TRUNCATE
CODING COST LOOKUP TABLE
CALCULATE LAMBDA/2
70

INTERNATIONAL SEARCH REPORT

Int[illegible]al Application No
PCT/GB 98/00582

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 H04N7/30

According to International Patent Classification(IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 H04N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EP 0 509 576 A (AMPEX) 21 October 1992 see abstract see page 3, line 17 - page 8, line 39; figures	1-17
A	EP 0 513 520 A (IBM) 19 November 1992 see abstract; claims; figures	1-17
A	EP 0 478 230 A (AMERICAN TELEPHONE & TELEGRAPH) 1 April 1992 see abstract see figures	1-17

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
14 May 1998	25/05/1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Gries, T

3

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 95 35628 A (SNELL & WILCOX LTD ;KNEE MICHAEL JAMES (GB); DEVLIN BRUCE FAIRBAIR) 28 December 1995 cited in the application see abstract see page 3, paragraph 2 - page 6, line 2	1-17
A	EP 0 711 079 A (RCA THOMSON LICENSING CORP) 8 May 1996 see the whole document	1-17
A	EP 0 599 258 A (MATSUSHITA ELECTRIC IND CO LTD) 1 June 1994 see abstract see column 3, line 20 - line 36	4,5,11, 12
A	EP 0 720 375 A (SONY CORP) 3 July 1996 see the whole document	1-17
A	EP 0 705 039 A (FUJI XEROX CO LTD) 3 April 1996 see abstract	1-17
A	US 5 521 643 A (YIM MYUNG-SIK) 28 May 1996 see abstract	1-17
A	EP 0 710 030 A (MITSUBISHI ELECTRIC CORP) 1 May 1996 see abstract see column 3, line 33 - column 5, line 49; figures	1-17
A	EP 0 739 138 A (AT & T CORP) 23 October 1996 see abstract see page 3, line 5 - line 14; figures 4-7	1-17
A	WO 96 34496 A (PHILIPS ELECTRONICS NV ;PHILIPS NORDEN AB (SE)) 31 October 1996 see abstract see page 2, line 21 - page 3, line 3; claims; figures	13
A	DE 35 11 659 A (SIEMENS AG) 2 October 1986 see abstract; figures	1,7,10

3

INTERNATIONAL SEARCH REPORT

Information on patent family members

In ternational Application No
PCT/GB 98/00582

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 0509576 A	21-10-1992	AT 162918 T	15-02-1998
		CA 2063972 A	19-10-1992
		DE 69224227 D	05-03-1998
		JP 5153406 A	18-06-1993
		MX 9201788 A	01-02-1993
EP 0513520 A	19-11-1992	US 5157488 A	20-10-1992
		CA 2062155 A	18-11-1992
		DE 69220541 D	31-07-1997
		DE 69220541 T	15-01-1998
		JP 5183758 A	23-07-1993
EP 0478230 A	01-04-1992	US 5038209 A	06-08-1991
		CA 2050102 A,C	28-03-1992
		DE 69124760 D	03-04-1997
		DE 69124760 T	12-06-1997
		JP 2509770 B	26-06-1996
		JP 4288776 A	13-10-1992
		KR 9504118 B	25-04-1995
WO 9535628 A	28-12-1995	AU 2744095 A	15-01-1996
		CA 2193109 A	28-12-1995
		EP 0765576 A	02-04-1997
		JP 10503895 T	07-04-1998
EP 0711079 A	08-05-1996	CN 1139351 A	01-01-1997
		JP 8214304 A	20-08-1996
EP 0599258 A	01-06-1994	JP 6165151 A	10-06-1994
		US 5568199 A	22-10-1996
EP 0720375 A	03-07-1996	JP 8256334 A	01-10-1996
		JP 8256064 A	01-10-1996
		US 5706009 A	06-01-1998
EP 0705039 A	03-04-1996	JP 8102943 A	16-04-1996
US 5521643 A	28-05-1996	NONE	
EP 0710030 A	01-05-1996	JP 8130743 A	21-05-1996

INTERNATIONAL SEARCH REPORT

Information on patent family members

Int ional Application No
PCT/GB 98/00582

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 0710030 A		AU 2052995 A	09-05-1996
		CA 2151085 A	01-05-1996
		CN 1126409 A	10-07-1996
		NO 952221 A	02-05-1996
		SG 33377 A	18-10-1996
EP 0739138 A	23-10-1996	CA 2173881 A	20-10-1996
		JP 8298464 A	12-11-1996
WO 9634496 A	31-10-1996	AU 5160996 A	18-11-1996
		EP 0768008 A	16-04-1997
		JP 10502792 T	10-03-1998
DE 3511659 A	02-10-1986	NONE	